Earth and Space Science

RESEARCH ARTICLE
10.1029/2019EA000997

Machine Learning Approach for Solar Wind Categorization

Hui Li[1,2], **Chi Wang**[1,3], **Cui Tu**[4,5], **and Fei Xu**[6]

[1]State Key 3 Laboratory of Space Weather, National Space Science Center, CAS, Beijing, China, [2]School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing, China, [3]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China, [4]Laboratory of Near Space Environment, National Space Science Center, CAS, Beijing, China, [5]College of Materials Sciences and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing, China, [6]Physics Department, Nanjing University of Information Science and Technology, Nanjing, China

Key Points:
- An eight-dimensional scheme for four-type solar wind categorization is developed based on 10 supervised machine learning classifiers
- Two application examples indicate that solar wind classification is useful for space weather early warning
- Although the accuracy drops by 1.5%, our scheme without composition information is good enough and has wide applicability

Correspondence to:
H. Li,
hli@nssc.ac.cn

Citation:
Li, H., Wang, C., Cui, T., & Xu, F. (2020). Machine learning approach for solar wind categorization. *Earth and Space Science*. *7*, e2019EA000997. https://doi.org/10.1029/2019EA000997

Received 10 NOV 2019
Accepted 23 MAR 2020
Accepted article online 14 APR 2020

Abstract Solar wind classification is conducive to understanding the ongoing physical processes at the Sun and in solar wind evolution in interplanetary space, and, furthermore, it is helpful for early warning of space weather events. With rapid developments in the field of artificial intelligence, machine learning approaches are increasingly being used for pattern recognition. In this study, an approach from machine learning perspectives is developed to automatically classify the solar wind at 1 AU into four types: coronal-hole-origin plasma, streamer-belt-origin plasma, sector-reversal-region plasma, and ejecta. By exhaustive enumeration, an eight-dimensional scheme (B_T, N_P, T_P, V_P, $N_{\alpha p}$, $T_{\exp}/T_P$, S_p, and M_f) is found to perform the best among 8,191 combinations of 13 solar wind parameters. Ten popular supervised machine learning models, namely, *k*-nearest neighbors (KNN), Support Vector Machines with linear and radial basic function kernels, Decision Tree, Random Forest, Adaptive Boosting, Neural Network, Gaussian Naive Bayes, Quadratic Discriminant Analysis, and eXtreme Gradient Boosting, are applied to the labeled solar wind data sets. Among them, KNN classifier obtains the highest overall classification accuracy, 92.8%. Although the accuracy can be improved by 1.5% when O^{7+}/O^{6+} information is additionally considered, our scheme without composition measurements is still good enough for solar wind classification. In addition, two application examples indicate that solar wind classification is helpful for the risk evaluation of predicted magnetic storms and surface charging of geosynchronous spacecraft.

1. Introduction

In 1959, solar wind observation was first made by the Soviet satellite, *Luna 1*. Since then, decades of in-situ solar wind measurements have firmly established that the solar wind plasma comes from different origins. Xu and Borovsky (2015) showed that the solar wind can generally be classified into four major types: coronal-hole-origin plasma (CHOP), streamer belt plasma (SBP), sector-reversal-region plasma (SRRP), and ejecta (EJECT).

CHOP is sometimes called the fast solar wind, which originates from the open-field line regions of coronal holes, and typically exhibits speeds in excess of 500 km/s at 1 AU and beyond (McComas et al., 2008; Sheeley et al., 1976). Statistically, CHOP tends to be homogeneous (Bame et al., 1977) with high proton temperature and low plasma density (Schwenn, 2006) and is dominated by outward propagating Alfvénic waves (Luttrell & Richter, 1988). It exhibits a statistical nonadiabatic heating of the protons between 0.3 and 1.0 AU (Hellinger et al., 2011). In addition, field-aligned relative drifts between the alpha particles and protons can frequently be found in CHOP, with a speed up to the local proton Alfvén speed (Marsch et al., 1982). Moreover, the relative fluctuations of magnetic field and solar wind velocity are large in CHOP, about 24% and 19%, respectively. However, the corresponding Fourier spectral indices are −1.56 and −1.55 (Borovsky, 2012), which is more likely due to the Iroshnikov-Kraichnan scaling ($f^{-3/2}$) for turbulence. As proposed by Li et al. (2011), this further indicates that current sheets are rare in such kind of solar wind.

SBP and SRRP are two subgroups of the streamer-belt-origin plasma (SBOP) (Antonucci et al., 2005; Schwenn, 2006, and the references therein), which is also known as the slow solar wind with a typical speed less than 400 km/s. Compared to CHOP, SBOP does not exhibit much Alfvénic fluctuations (Schwenn, 1990)

but is highly structured (Bame et al., 1977) with low proton temperature and high plasma density (Schwenn, 2006). In addition, the alpha-proton relative drift is typically absent in SBOP (Asbridge et al., 1976), and the protons are closer to adiabatic (Eyni & Steinitz, 1978). The relative fluctuations of magnetic field and solar wind velocity are small in SBOP, which are only 16% and 11%, respectively. Different from the situations in CHOP, both of the corresponding Fourier spectral indices obey Kolmogoroff's law ($f^{-5/3}$), giving −1.70 and −1.67, respectively (Borovsky, 2012). This indicates that the solar wind may contain many current sheet structures (Li et al., 2011). The origin mechanism of SBP at the Sun is still a major unsolved problem in solar physics. There are two main mechanisms of SBP origin: One is the interchange magnetic reconnection of open-field lines with closed streamer belt field lines (Antiochos et al., 2011; Crooker et al., 2012; Fisk et al., 1999; Subramanian et al., 2010); the other one is from the edge of a coronal hole near a streamer belt (Arge et al., 2003; Wang & Sheeley, 1990). SRRP is suggested to be emitted from the top of the helmet streamers (Foullon et al., 2011; Gosling et al., 2012; Suess et al., 2009). Statistically, SBP and SRRP have different characteristics in the solar wind and subsequent effects on the geospace environments, which have been summarized by Borovsky and Denton (2013).

Another major category of solar wind plasma is the so-called EJECT, which is associated with solar transients such as interplanetary coronal mass ejections (ICMEs) and magnetic clouds (MCs) (Richardson et al., 2000; Zhao et al., 2009). The origination of EJECT is the magnetic reconnection associated with the structures of streamer belts or active regions, which can impulsively emit plasma and make the magnetic field deviate from the Parker spiral (Borovsky, 2010). The typical signatures of EJECT at 1 AU have been well summarized (see Zurbuchen & Richardson, 2006, and the references therein), for example, enhanced and smoothly rotating magnetic field, low proton temperature and plasma β, extreme density decrease, enhanced density ratio between alpha and proton, abundance and charge state anomalies of heavy ion species, bidirectional strahl electron beams, cosmic ray depletion, and declining velocity. Different from the expansions of CHOP or SBOP in the two directions transverse to radially outward from the Sun, impulsive EJECT expands in all three directions as they propagate outward (Klein & Burlaga, 1982). Recently, Li et al. (2016) performed a statistical survey on Alfvénic fluctuations inside ICMEs, finding that only 12.6% of EJECT are found to be Alfvénic, and such a percentage decays linearly in general as the radial distance increases. The relative fluctuations of magnetic field and solar wind velocity are medium in EJECT, 21% and 15%, respectively (Borovsky, 2012). The Fourier spectral indices are close to −5/3 (Borovsky, 2012) and may decrease with the radial distance (Li et al., 2017).

The categorization of solar wind into its origin is of great importance for solar and heliospheric physics studies. First, the statistical properties of solar wind should be clarified by its type to give a more comprehensive understanding. Second, dividing the solar wind observations at 1 AU according to their origins can lead to a better diagnosis of physical processes ongoing at the Sun (Borovsky, 2008; Mariani et al., 1983; Matthaeus et al., 2007; Thieme et al., 1989, 1990; Zastenker et al., 2014). Third, the geoeffectiveness (geomagnetic activity, specifically, magnetic storm and substorm) of solar wind from different origins varies considerably (e.g., Borovsky & Denton, 2006, 2013; Turner et al., 2009). Such a categorization would be helpful for space weather early warning. Note that these differences are in statistical terms. For individual cases, the situations may be quite different and complicated.

Usually, solar wind classification is done by experienced people. In the literature, several empirical categorization methodologies in different parameter spaces have been proposed. In a one-dimensional parameter space, the solar wind was usually separated into "fast wind" or "slow wind" according to its speed, V_p (Arya & Freeman, 2012; Feldman et al., 2005; Tu & Marsch, 1995; Yordanova et al., 2009). However, such a V_p scheme can only roughly divide the solar wind into CHOP and SBOP but could not separate out EJECT, SBP, and SRRP. Moreover, the criterion of V_p is not unique. In 2014, another one-dimensional scheme based on the parameter P_{type} ($= 2\log S_p - \log(C^{6+}/C^{5+}) - \log(C^{7+}/C^{6+})$) was proposed by Borovsky and Denton (2014). With better understanding of ICMEs and MCs, many methodologies have been proposed to identify EJECT (see Kunow et al., 2006; Zurbuchen & Richardson, 2006, and the references therein), and several catalogs of EJECT at 1 AU have been produced (e.g., Jian et al., 2006; Lepping et al., 2005; Richardson & Cane, 2010). Recently, the composition measurements were used for solar wind classification. An algorithm in a two-dimensional parameter space, such as O^{7+}/O^{6+} and V_p, was constructed by Zurbuchen et al. (2002), Zhao et al. (2009), and von Steiger et al. (2010). Such a two-dimensional scheme is still not able to divide SBOP into SBP and SRRP. In addition, such a scheme is not generally available for most solar wind measuring spacecraft due to the lack of onboard ion composition instruments. Xu and Borovsky (2015)

developed a three-parameter, four-plasma-type categorization scheme based on the commonly used solar wind measurements and obtained a good classification accuracy. In addition, an onboard solar wind classification algorithm was already applied in the *Genesis* spacecraft (Neugebauer et al., 2003; Reisenfeld et al., 2003). Such an automatic method requires the measurement of bidirectional electron and historic solar wind classification results.

Although the traditional classification has significant improvements in recent decades, there remains some room for improvement for the existing empirical categorization schemes. The multilabel classification is regarded as a typical task of machine learning. Recently, the performance of machine learning classification is getting much better as a result of the rapid developments of artificial intelligence theory and techniques. Machine learning techniques are becoming more and more popular and powerful in big-volume data analysis in space physics, which may offer a solution to improve the accuracy of solar wind classification. As a pioneer, Camporeale et al. (2017) recently employed a machine learning technique, Gaussian process (GP), in a four-category solar wind classification, and obtained a median accuracy larger than 96.0% for all categories. However, the time resolutions of the variables they used are not uniform. For example, the temporal resolution is 1 day for sunspot number and solar radio flux (10.7 cm) but is 1 hr for the other five solar wind parameters and for the reference solar wind data sets. Camporeale et al. (2017) did not demonstrate the reasonableness of such mixture of hourly averaged solar wind parameters and daily sampled parameters.

To further demonstrate the application of machine learning techniques in solar wind classification, 10 additional popular supervised machine learning models are applied to classify the solar wind plasma into four plasma types (CHOP, SBP, SRRP, and EJECT) in this work. To expand the application scope of our classification scheme, only some typical solar wind observations with the same temporal resolution are used, such as N_α, N_p, B_T, T_p, and V_p. In addition, two examples of solar wind classification are applied to the risk evaluation of predicted magnetic storms and surface charging of geosynchronous spacecraft.

2. Methodology

For conventional classifications of the solar wind plasma at 1 AU, reference solar wind data with known plasma types should be first collected. Then, empirical relationships are developed to describe the domains of different plasma in some parameter space. In general, human intuition performs well in two- and three-dimensional parameter space but cannot easily derive empirical relationships in a multidimensional space.

For supervised machine learning approaches, reference solar wind data with known plasma types are needed for training the classifier as well. Then, the discriminant rules would be developed automatically by machine learning classifiers. The advantage is that the discriminant rules can be easily obtained in a multidimensional space for the machine learning perspective. Usually, 75% (80%) of the reference solar wind data are used for training, and the remaining 25% (20%) are used for testing, especially for the situation with the cases less than 10,000. Sometimes, a validation set is recommended to tune the parameters of a classifier, and its ratio depends on both the level of accuracy and the standard error.

2.1. Machine Learning Classifiers

Classification is regarded as one of the typical tasks carried out by so-called machine learning system. The classifier is a critically important part of machine learning toolkit. As a result of the rapid development of machine learning technique, a large number of classification algorithms have been developed. In this study, we will apply 10 widely used classifiers (Cady, 2017) to perform solar wind categorization, namely, *k*-nearest neighbors (KNN), linear support vector machine (LSVM), Support Vector Machine with a kernel of Gaussian Radial Basis Function (RBFSVM), Decision Tree (DT), Random Rorest (RF), Adaptive Boosting (AdaBoost), Neural Network (NN), Gaussian Naive Bayes (GNB), Quadratic Discriminant Analysis (QDA), and eXtreme Gradient Boosting (XGBoost). Table 1 gives the references of these 10 classifiers for readers to get more details. All the classification algorithms are included in the Scikit-learn package, which is an open-source machine learning library written in the Python programming language (Pedregosa et al., 2011). In this work, we will use the Scikit-learn package to carry out solar wind classifications. The details of the scikit-learn package can be found at the website (http://scikit-learn.org/stable/index.html).

Table 1
Ten Machine Learning Classifiers Used in This Study

Classifier	Abbreviation	Reference
k-nearest neighbors	KNN	Denoeux (1995)
Linear support vector machine	LSVM	Fan et al. (2008)
SVM with Gaussian radial basis function kernel	RBFSVM	Buhmann (2003)
Decision tree	DT	Breiman et al. (1984)
Random forest	RF	Ho (1995)
Adaptive Boosting	AdaBoost	Zhu et al. (2009)
Neural network	NN	Rojas (1996)
Gaussian naive Bayes	GNB	Perez et al. (2006)
Quadratic discriminant analysis	QDA	Srivastava et al. (2007)
eXtreme Gradient Boosting	XGBoost	Chen and Guestrin (2016)

2.2. Reference Solar Wind Data

For supervised machine learning, reference solar wind data sets with known types are needed to train the classifiers. We use the same data sets utilized in Xu and Borovsky (2015), and the solar wind plasma will be divided into four types: CHOP, SBP, SRRP, and EJECT. The collection of reference CHOP comes from the ideal events used by Xu and Borovsky (2015). They examined the solar wind speed V_p, the proton-specific entropy $S_p = T_p/N_p^{2/3}$, O^{7+}/O^{6+}, C^{6+}/C^{5+}, and the characteristics of the interplanetary magnetic field to identify CHOP. The intervals of 27-day repeating steady high-speed solar wind streams with long intervals (days) are regarded as CHOP. CHOP starts after the compression of the corotating interaction region (CIR) and ends before the onset of the trailing edge rarefaction. At the same time, they also excluded large jumps in S_p, O^{7+}/O^{6+}, or C^{6+}/C^{5+} to make sure CHOP was not contaminated with EJECT. A total of 3,049 hr of CHOP identified by Xu and Borovsky (2015) is used here.

The collection of reference SBP comes from the pseudostreamers during 2002–2008 identified by Borovsky and Denton (2013). Looking earlier in time the plasma upstream of the CIR, they checked the preceding intervals of CHOP. If the preceding coronal hole was of the same magnetic sector as the coronal hole immediately following the CIR, and if no sector reversals occurred in the SBOP between the successive two coronal holes, then the SBOP was classified into SBP. A total of 2,275 hr of SBP identified by Borovsky and Denton (2013) is used here.

The collection of reference SRRP also comes from the work done by Xu and Borovsky (2015). They examined the electron strahl observation and found some broad regions where the electron strahl dropped out around magnetic sector reversals at 1 AU. They denoted the regions where the strahl was very weak, intermittent, and/or intermittently bidirection just outside the strahl dropped out regions, to be "strahl confusion zones." The solar wind from these confusion zones is defined as SRRP. A total of 1,740 hr of SRRP is used here.

The MC collection made by Lepping et al. (2005) is used to represent EJECT here, which can be found at the website (https://wind.gsfc.nasa.gov/mfi/mag_cloud_pub1.html). MCs are believed to be a subset of ICMEs with an enhancement of magnetic field intensity and a gradual rotation in direction. The typical properties of MC are a flux rope field configuration, low proton temperatures, and low plasma beta value (Klein & Burlaga, 1982). In general, only about one third of ICMEs can be regarded as MCs (Bothmer & Schwenn, 1996; Richardson & Cane, 2004). Xu and Borovsky (2015) found a dual-population structure for the collection of ICMEs identified by Richardson and Cane (2010), but a single population for the collection of MCs identified by Lepping et al. (2005). They believed that MCs can better present EJECT from the Sun, while the collection of ICMEs probably contains some non-EJECT data. A total of 1,926 hr of EJECT is used here.

After removing some data gaps, the reference data set is composed of 2,881 (33.4%) 1-hr events categorized as CHOP, 2,215 (25.7%) events of SBP, 1,694 (19.6%) events of SRRP, and 1,835 (21.3%) events of EJECT. The imbalance ratio of these four types of solar wind may affect the classification accuracy. In general, the accuracy would be relatively low when fewer reference solar wind data are used for training. The ratio of reference SRRP is the lowest. Its classification accuracy is indeed found to be lower than the other three types in the following section. The solar wind parameters used in this study are from the low-resolution, hourly

Table 2
List of 13 Parameters Used for Solar Wind Classification

Parameter	Symbol
Magnetic filed intensity	B_T
Proton density	N_p
Proton temperature	T_p
Solar wind speed	V_p
Proton-specific entropy	S_p
Alfveń speed	V_A
Temperature ratio	$T_{\exp}/T_p$
Ratio of proton and alpha number density	$N_{\alpha p}$
Dynamic pressure	P_d
Solar wind electric field	E_y
Plasma beta value	β
Alfvén Mach number	M_A
Fast magnetosonic Mach number	M_f

averaged data from the OMNI database (https://omniweb.gsfc.nasa.gov/), a 1963-to-current compilation of near-Earth solar wind magnetic field and plasma parameter data compiled from several spacecraft in geocentric or L1 (Lagrange point) orbits.

3. Categorization Results

With the input of solar wind parameters and information of solar wind types, the classifiers can build discriminant rules automatically based on machine learning algorithms. Note that most solar wind measuring spacecraft have no composition instrumentation. To make the applicability of our classification scheme more extensive, the typical solar wind observations (the magnetic field intensity, B_T, the proton number density, N_p, the alpha particle number density, N_α, the proton temperature, T_p, and the solar wind speed, V_p) and their derived quantities are used here. As listed in Table 2, a total of 13 parameters are used for solar wind classification, such as B_T, N_p, T_p, and V_p, the proton-specific entropy, S_p, the Alfveń speed, $V_A = B_T/\sqrt{\mu_0 m_p N_p}$ (μ_0 is the permeability in vacuum and m_p is the mass of proton), the temperature ratio, $T_{\exp}/T_p$ ($T_{\exp} = (V_p/258)^{3.113}$ is the velocity-dependent expected proton temperature given by Xu and Borovsky (2015) in unit of eV), the number density ratio of proton and alpha, $N_{\alpha p}$, the dynamic pressure, P_d, the solar wind electric field, E_y, the plasma beta value β, the Alfvén Mach number, $M_A = V_A/V_p$, and the fast magnetosonic Mach number, $M_f = V_p/\sqrt{C_s^2 + V_A^2}$ (C_s is the acoustic velocity). The electron temperature is assumed to be 1.4×10^5 K based on 1978–1982 ISEE-3 data (Newbury et al., 1998). Note that this parameter list includes all the parameters used in Xu and Borovsky (2015) and four of seven parameters used in Camporeale et al. (2017). As mentioned, the reference solar wind with known types is from the hourly averaged OMNI database; thus, only the parameters with a temporal resolution of 1 hr are considered here. The parameters with a temporal resolution of 1 day used in Camporeale et al. (2017), such as the sunspot number and solar radio flux (10.7 cm), are not considered here. Among them, a specific combination of parameter with the highest classification accuracy will be chosen for further analysis.

Figure 1 shows the probability density distributions of the above 13 parameters calculated from the whole reference solar wind data sets. Similar probability density distributions of V_p, V_A, S_p, and $T_{\exp}/T_p$ are also shown by Camporeale et al. (2017). Note that the parameters have been rescaled as follows: $X = (X - \bar{X})/\sigma_X$, where $\bar{X}$ represents the mean value of a parameter and σ_X denotes the standard deviation. Obviously, it is difficult to distinguish the four-type solar wind well from any individual probability distribution, which motivates the classification in a multidimensional space. Nevertheless, some parameters could contribute to distinguish some solar wind type from the others. For example, B_T and M_f contribute to distinguish EJECT from the others, especially from the SRRP; N_p, V_p, and $N_{\alpha p}$ are useful to distinguish between CHOP and SRRP; T_p and S_p help to distinguish CHOP from the others; and V_A is helpful to distinguish SRRP from the others. A natural thought is that the classification accuracy would be improved greatly by considering

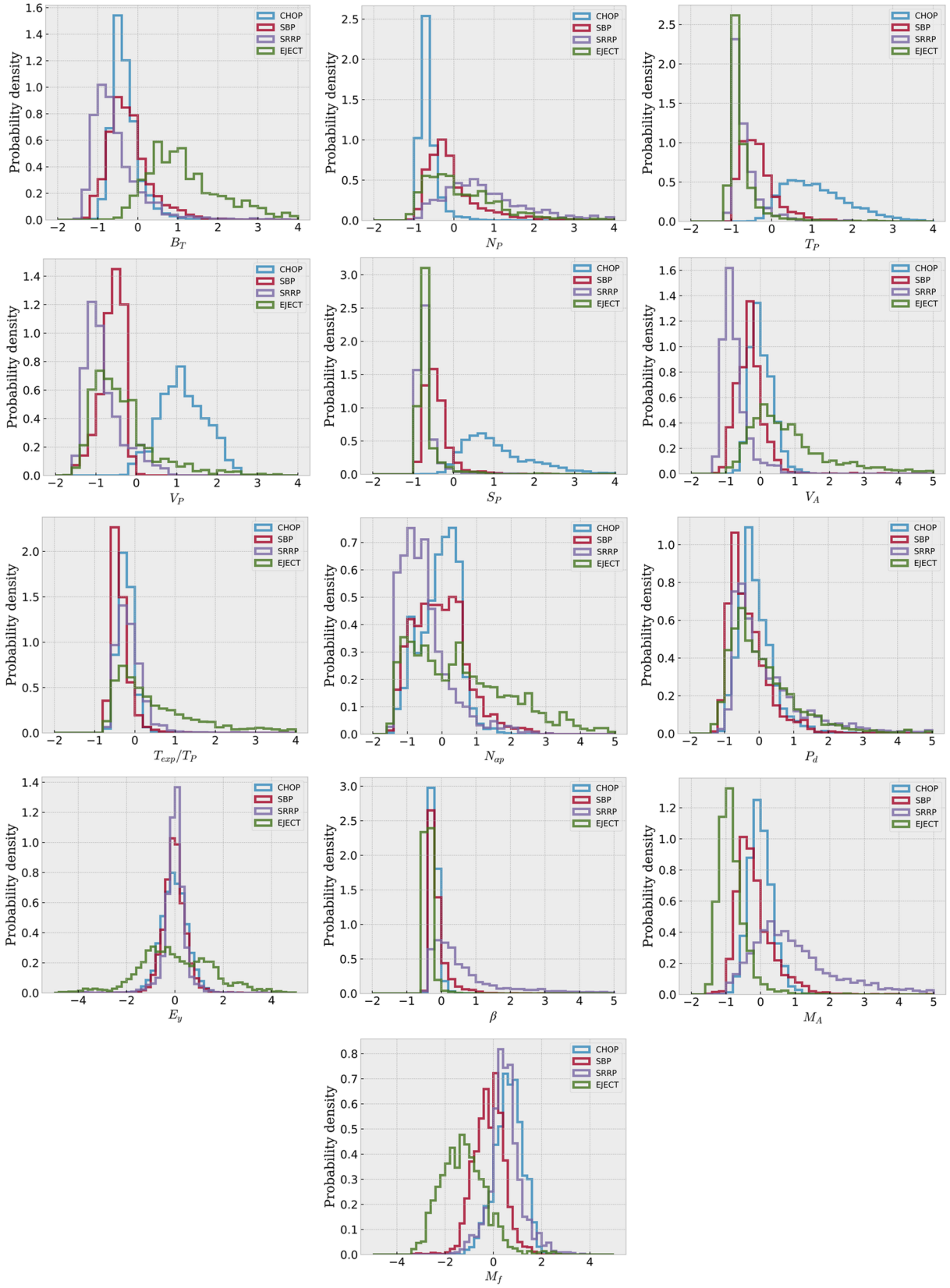

Figure 1. Probability density distributions of 13 solar wind parameters calculated from the whole reference solar wind data sets. The parameters have been rescaled as follows: $X = (X - \bar{X})/\sigma_X$. The area under each curve is equal to 1.

Table 3
Classification Performances for 10 Classifiers Based on the Combination of B_T, N_p, T_p, V_p, $N_{\alpha p}$, T_{exp}/T_p, S_p, and M_f

	CHOP	SBP	SRRP	EJECT	4-type	HKSS
KNN	99.2	91.1	83.8	92.9	92.8	0.902
XGBoost	99.2	90.9	83.6	92.8	92.6	0.898
RF	99.3	90.2	81.6	94.1	92.3	0.895
RBFSVM	99.1	89.0	81.1	94.1	91.9	0.890
NN	99.1	88.7	80.6	92.2	91.3	0.881
DT	98.1	84.8	77.6	89.0	88.7	0.846
LSVM	99.0	81.1	71.1	88.2	86.6	0.816
QDA	98.7	80.4	75.0	73.7	84.0	0.779
GNB	96.8	76.0	76.9	73.1	82.5	0.767
AdaBoost	97.5	85.1	45.2	85.6	81.5	0.737

Note. From the second to sixth columns, the value gives the classification accuracy. The last column gives the Hanssen-Kuiper Skill score (HKSS). Note that 75% of the reference solar wind data are used for training and the remaining 25% are used for testing. The training data set is randomly selected by 100 times, and the mean accuracies are given here.

the above eight parameters together. Actually, the selected eight-dimensional parameter scheme with the best classification accuracy for KNN classifier contains seven of the above eight parameters, only with V_A replaced by $T_{\rm exp}/T_p$.

Given 13 input features, a total of 8,191 combinations exists. Taking the KNN classifier as an example, the classification accuracy is calculated by using all the 8,191 combinations of input features. In the eight-dimensional scheme, the combination of B_T, N_P, T_P, V_P, $N_{\alpha p}$, $T_{\rm exp}/T_P$, S_p, and M_f is found to perform the best, with the overall accuracy of 92.8%. The accuracy for classifying CHOP, SBP, SRRP, and EJECT is 99.2%, 91.1%, 83.8%, and 92.9%, respectively. Although this scheme is selecting from 8,191 combinations of 13 variables for systemics, it has a physical interpretation. As shown in Figure 1, these parameters indeed contribute to distinguish some solar wind type from the others. If some new variables are considered, another method to determine the variable combination may also work and reduce the test number greatly. For example, identify the first variable, by using that alone the best classification accuracy can be obtained. Then, identify the second variable, by considering that with the first determined variable together, the best classification accuracy can be obtained. Finally, repeat the second step until the accuracy cannot be improved by adding any new variable. Actually, a set of mutually independent variables contain enough information of the classification system. Here, some combined parameters, for example, S_p, V_A, and $T_{\rm exp}/T_p$, are used only for the purpose of improving the classification accuracy. If the mutually independent variables (B_T-V_p-N_p-T_p-$N_{\alpha p}$) are used, the classification accuracy of the four-type solar wind will decline slightly from 92.8% to 92.0%. For the KNN classifier, the number of neighbors is set to be four. The weighting scheme is chose to be "distance", which means that closer neighbors of a query point will have a greater influence than neighbors that are farther away. And the standard Euclidean metric is used here.

The classification is also done for the other nine classifiers with the same parameter scheme used. The results are listed in Table 3. Five classifiers, KNN, XGBoost, RF, RBFSVM, and NN, produce an accuracy better than 90%. DT and LSVM also perform well, with the overall accuracy better than 85%. The remaining classifiers, QDA, GNB, and AdaBoost, yield accuracies between 80% and 85%. It should be mentioned that the overall accuracy of the other nine classifiers could be improved if some special kind of parameter combination was to be used. The identification of CHOP is relatively easy. All 10 classifiers work very well, with an accuracy better than 96.5% and the highest accuracy given by RF of 99.3%. For identifying EJECT, the accuracy decreases slightly. Only five classifiers yield accuracies better than 92%, and the highest accuracy given by RBFSVM is 94.1%. For identifying SBP, only three classifiers yield accuracies better than 90%, with the highest accuracy given by KNN of 91.1%. The identification of SRRP is relatively difficult. Only five classifiers yield accuracies better than 80%, and the highest accuracy given by KNN is only 83.8%. Note that 75% of the reference solar wind data are used for training and the remaining 25% are used for testing. To make

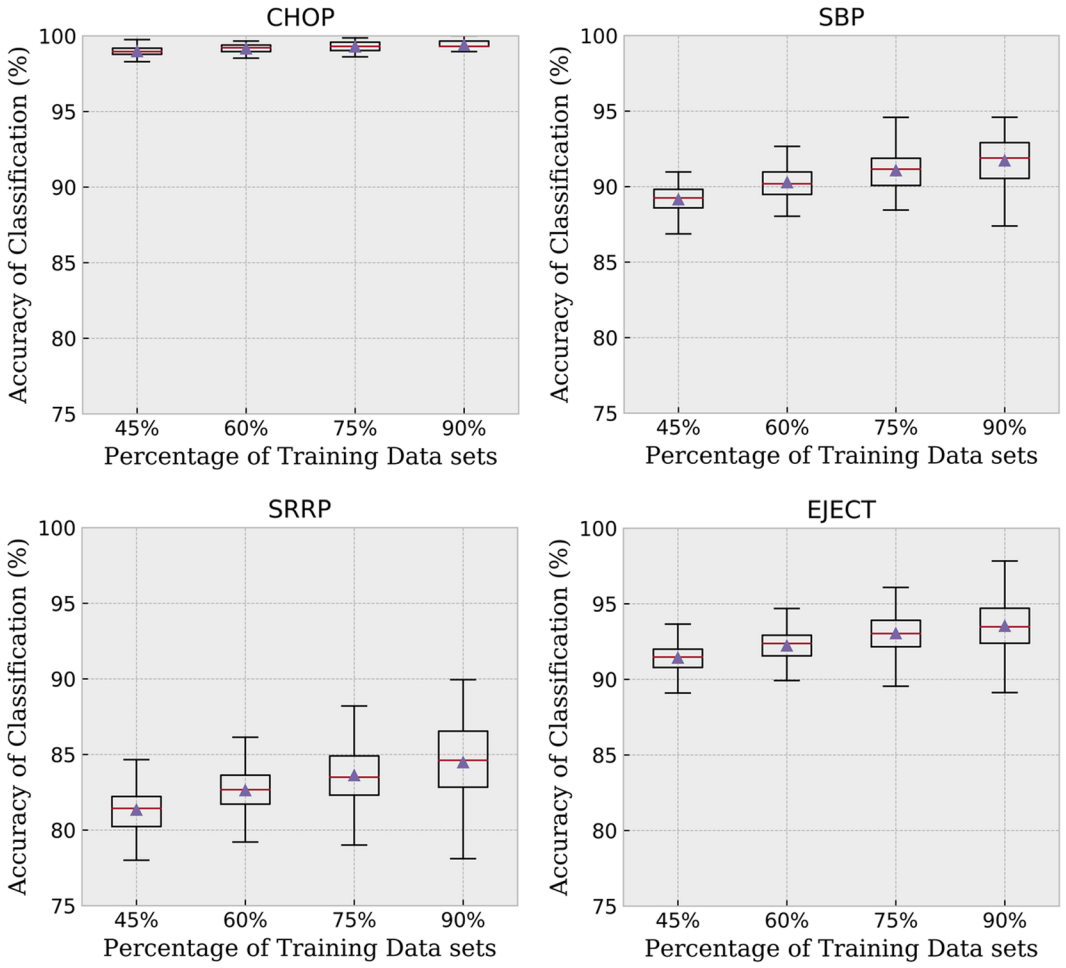

Figure 2. Accuracy of the KNN classifier calculated from 100 runs with different ratio of training data set being chosen randomly. The boxes denote the first and third quartiles of the accuracy distribution. The horizontal lines and triangles represent the median and mean values, respectively. The whiskers denote the 2nd and 98th percentiles.

sure that our results are independent on the choice of training data set, cross-validation is quite necessary. Thus, we randomly select the training data set by 100 times. The accuracy given in Table 3 is the averaged value of these 100 tests.

In addition to the classification accuracy, the Hanssen-Kuiper skill score, HKSS, or Hanssen-Kuiper discriminant, is also given in Table 3. The HKSS, also known as the true skill statistic or Pierce skill score, represents the classification accuracy relative to that of random chance. For multicategory classification, its expression can be written as follows:

$$HKSS = \frac{\frac{1}{N}\sum_{i=1}^{K} n(F_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \sum_{i=1}^{K}(N(O_i))^2}, \quad (1)$$

where $n(F_i, O_i)$ denotes the number of classifications in category i that had observations in category i, $N(F_i)$ denotes the total number of classifications in category i, $N(O_i)$ denotes the total number of observations in category i, and N is the total number of classification. HKSS ranges from −1 to 1. 1 represent the perfect performance, 0 denotes no improvement over a reference classification, and ≤0 indicates worse than the reference. From Table 3, it is clear that the results of HKSS for the 10 classifiers have a similar trend with the results of accuracy.

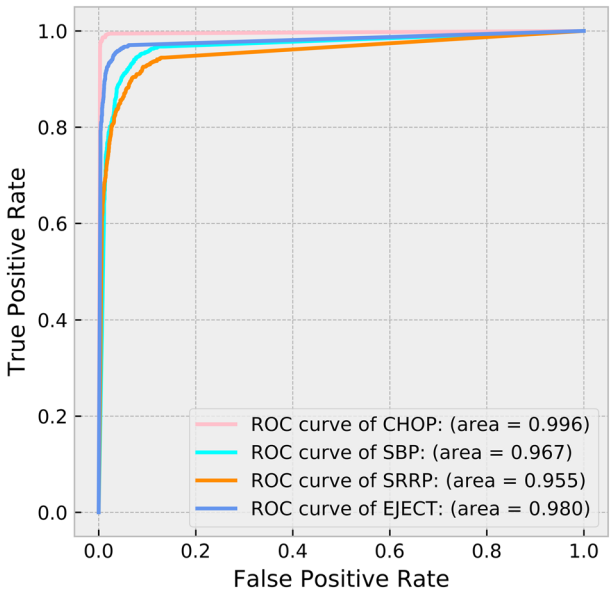

Figure 3. Receiver operating characteristic (ROC) curves for CHOP, SBP, SRRP, and EJECT. The false positive rate is defined as the ratio of false positives divided by the total number of negatives. The true positive rate denotes the ratio of true positives divided by the total numbers of positives. The area of the curve represents the goodness of binary classification, and unity denotes the perfect result.

To test the sensitivity to an individual variable in our eight-dimensional scheme, one variable is in turn left out from the scheme, and the accuracies are recalculated accordingly. When S_p is not considered, the classification accuracy has the least decrease, 0.1%. And the accuracy has the largest decrease, 2.2%, when $N_{\alpha p}$ is not considered. However, this does not imply that S_p is the least important variable in solar wind classification. Actually, for a one-parameter scheme, the highest classification accuracy is obtained by using S_p alone, among the 13 variables. For different parameter combination, the most sensitive parameter could be different as well.

It is hard to make sure that the result of supervised machine learning is neither overfitted nor underfitted. Comparing the accuracy of training versus testing data sets is a good way, but not sufficient. Cross-validation is another strategy to overcome such problems. Following the methodology of Camporeale et al. (2017), we also compare the results of 100 runs for different ratios of the training data. In general, overfitting is especially likely in cases where training examples are rare. Thus, a relative large ratio of training data, for example, 45%, 60%, 75%, and 90%, is used, and the results are shown in Figure 2. The boxes denote the first and third quartiles of the accuracy distribution. The horizontal lines and triangles represent the median and mean values, respectively. The whiskers denote the 2nd and 98th percentiles. It is clear that the mean accuracy slightly increases when the ratio of training data increases from 45% to 75%. For the ratio of 90%, the accuracy has no significant improvement; however, the variation amplitude of classification accuracy increases, and the lowest accuracy even slightly decreases slightly for identifying SBP, SRRP, and EJECT. In the following texts, the accuracies are all obtained by using 75% of the data for training. This is just a simple approach to judge whether an overfitting occurs or not. There may exist other, more robust, means of examining overfitting or underfitting. Camporeale et al. (2017) showed the accuracy of the GP classification model with 10%, 15%, 20%, and 25% of the original data used for training. Similarly, the accuracy increases when more data are used for training.

For binary classification, the probability threshold changes to accuracy in terms of true and false positives and negatives. Here, “true/false” denotes correct, or incorrect, classification, and “positive/negative” denotes that the solar wind is classified to be, not to be, some type. Thus, “true positive/false positive” denotes that the solar wind is correctly/incorrectly classified to be some type, while “true negative/false negative” denotes that the solar wind is correctly/incorrectly classified not to be some type. The receiver operating characteristic (ROC) curve for different values of thresholds gives a concise representation of this metric. The horizontal axis is the false positive rate (FTR), which is defined as the ratio of false positives

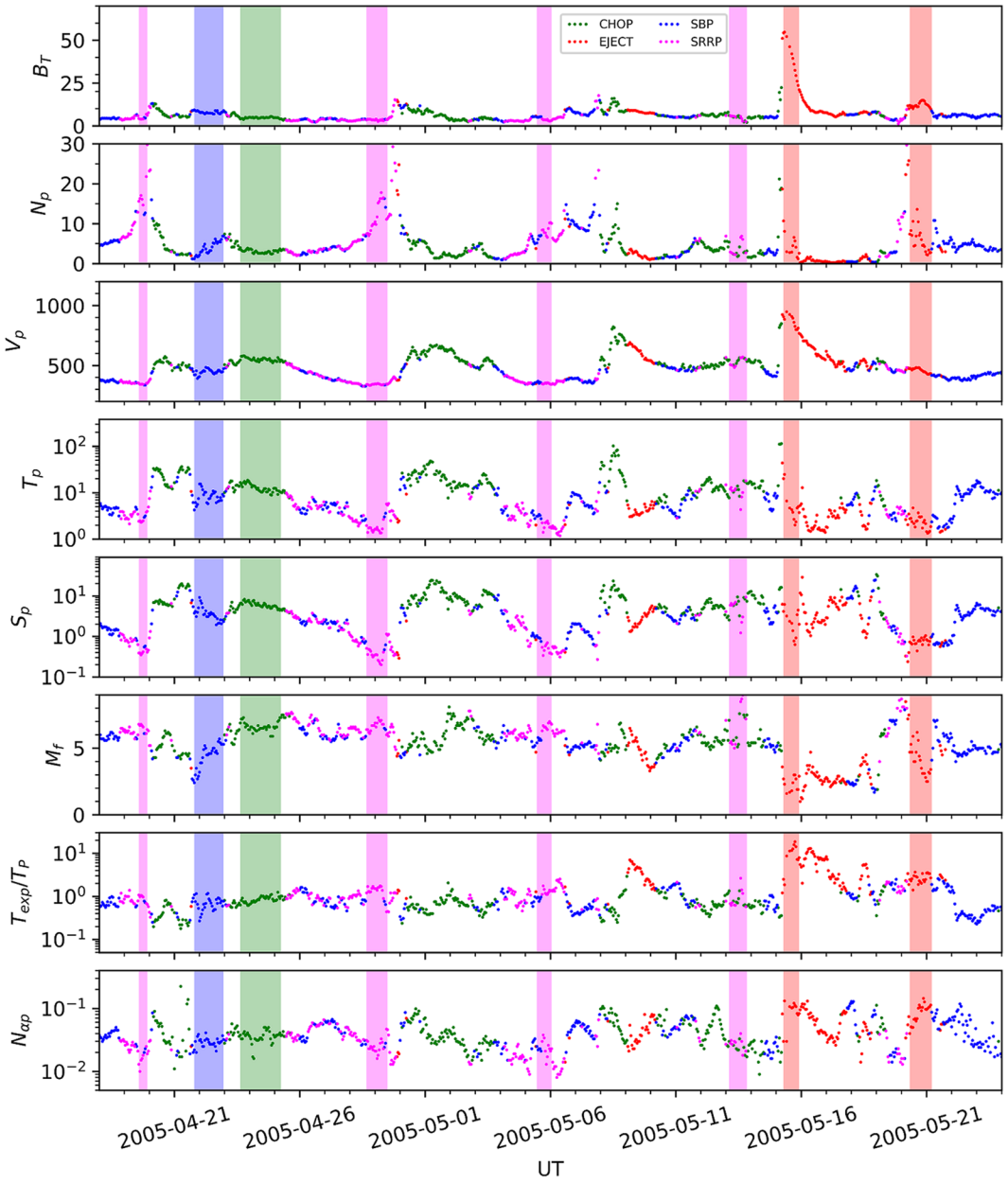

Figure 4. An example of solar wind classification obtained by the KNN classifier. From top to bottom, the panel represents the magnetic field intensity, the proton number density, the solar wind speed, the proton temperature, the proton-specific entropy, the plasma beta value, the fast magnetosonic Mach number, the dynamic pressure, and the ratio of proton and alpha number density. The units are in nT, cm^{-3}, km/s, eV, eV cm^2, unity, unity, and unity, respectively. The shaded regions represent the time intervals of reference solar wind with known types.

divided by the total number of negatives. And the vertical axis is the true positive rate (TPR), which denotes the ratio of true positives divided by the total numbers of positives. A perfect classification would give FPR = 0, TPR = 1, and the area of ROC curve equals unity. Figure 3 shows the ROC curves for CHOP, SBP, SRRP, and EJECT. The areas of the curves are 0.996, 0.967, 0.955, and 0.980, respectively, indicating that the classification is pretty good. For practice, the probability threshold can be chosen to be 0.3–0.5 to obtain optimal FPR and TPR, which is consistent with Camporeale et al. (2017).

Figure 4 shows an example of solar wind classification obtained by the KNN classifier. The shaded regions represent the time intervals of reference solar wind with known types. In general, all the solar wind can be distinguished well. It is clear that the CHOP, SBP, and EJECT in the shaded regions are identified perfectly

Table 4
Accuracies of Various Categorization Schemes in Solar Wind Classification

Accuracy (%)	CHOP	SBP	SRRP	EJECT	4-type
O^{7+}/O^{6+}-V_p	98.0	73.0		63.5	
Zhao et al. (2009)					
S_p-V_A-T_{exp}/T_p	96.9	69.9	72.0	87.5	83.2
Xu and Borovsky (2015)					
S_p-V_A-T_{exp}/T_p	97.2	74.9	69.7	88.7	84.3
KNN (this work)					
V_p-σ_T-SSN-$F10.7$-V_A-S_p-T_{exp}/T_p	99.7	98.7	97.5	96.1	98.2
Camporeale et al. (2017)					
V_p-σ_T-SSN-$F10.7$-V_A-S_p-T_{exp}/T_p	99.6	95.2	88.5	93.0	94.9
KNN (this work)					
B_T-N_P-T_P-V_P-$N_{\alpha p}$-T_{exp}/T_P-S_p-M_f	99.2	91.0	84.1	92.0	92.6
GP (isotropic Gaussian kernel)					
B_T-N_P-T_P-V_P-$N_{\alpha p}$-T_{exp}/T_P-S_p-M_f	99.2	91.1	83.8	92.9	92.8
KNN (this work)					

Note. Note that 25% of the database is used for training in Camporeale et al. (2017), but the ratio is 75% in our study. The training data set is randomly selected by 100 times, and the mean accuracies are given here.

with the accuracy nearly 100%. The classification accuracy for SRRP is not so high but still good, ~85%. Occasionally, it is wrongly identified as SBP (on 19 and 28–29 April and 5 May) or CHOP (on 13 May). Two long-duration EJECTs are also identified after CHOPs, for example, the EJECT on 15–17 and 20 May, which had already been identified as two MCs by Lepping et al. (2005). At the same time, some short-duration EJECTs (several hours) are also identified on 9–10 and 30–31 May, which may be the so-called small flux ropes proposed by Moldwin et al. (2000), and are in agreement with the small-scale magnetic flux rope database (https://fluxrope.info/) given by Dr. Jinlei Zheng and Dr. Qiang Hu at the University of Alabama in Huntsville. This indicates that our categorization scheme may in certain cases be useful for identifying small flux ropes, but more investigation and validation are needed.

Table 4 gives the comparison of the performances of various categorization schemes. The O^{7+}/O^{6+}-V_p scheme proposed by Zhao et al. (2009) cannot distinguish SBP and SRRP and does not work well for identifying EJECT. The accuracy is only 63.5%. Xu and Borovsky (2015) proposed the S_p-V_A-T_{exp}/T_p scheme, which has a significant improvement on identifying EJECT and increases the accuracy to 87.5%. In addition, such a scheme can also distinguish SBP and SRRP, with an accuracy ~70%. Note that the classification accuracies obtained by Xu and Borovsky (2015) are quite comparable to those obtained by KNN classifier. By taking the advantage of machine learning on classification in multidimensional parameter space, we apply an eight-dimensional scheme, the B_T-N_P-T_P-V_P-$N_{\alpha p}$-T_{exp}/T_P-S_p-M_f scheme, on KNN classifier, and obtain significant improvements in classification accuracies. The improvements of accuracy for identifying CHOP, SBP, SRRP, and EJECT are 2.3%, 21.2%, 11.8%, and 5.4%, respectively. For the four-type solar wind classification, the overall accuracy has an improvement of 9.6%. It should be mentioned that the feature space has been optimized only for the KNN approach. For other classifiers with some other parameter scheme used, the accuracies could be improved. Camporeale et al. (2017) proposed a classification scheme based on GP classifier with a combination of an isotropic Gaussian and a piecewise polynomial kernel with compact support. With the choice of the V_p-σ_T-SSN-$F10.7$-V_A-S_p-T_{exp}/T_p scheme (where σ_T is the standard deviation of proton temperature, SSN is the sunspot number, and $F10.7$ is the solar radio flux at 10.7 cm), the authors obtained classification accuracies better than 96% in all four solar wind category types. Note that 25% of the database is used for training in Camporeale et al. (2017), but the ratio is 75% in our study. If the same parameter scheme was performed on KNN classifier, the overall classification accuracy has a slight decrease of 3.3%, although the accuracy for SRRP has a decrease of 9.0%, as listed in Table 4. This indicates that the performance of KNN classifier is close to, or not far worse than, that of GP classifier used by Camporeale et al.

(2017). For comparison, the same eight-dimensional scheme is applied on the GP classifier with an isotropic Gaussian kernel. The classification accuracies are comparable to those obtained by the KNN classifier.

We apply the trained KNN classifier to classify the OMNI data set from 1963 to 2017. The probabilities of CHOP, SBP, SRRP, and EJECT are obtained. As mentioned before, the threshold of probability is chosen to be 0.3–0.5 to obtain optimal TPR and FPR. The event with the maximum probability less than the threshold is defined as an "undecided" event. If the threshold is chosen to be 0.3, the percentage of "undecided" events is 0.02%. And if the threshold is chosen to be 0.5, the percentage of "undecided" events is less than 2.2%. For comparison, the percentage of "undecided" events is 0.2% and 7.5% in Camporeale et al. (2017), indicating that the possibility of "undecided" solar wind type is larger than our approach.

4. Discussion

4.1. Daily Sampled Parameters Are Not Recommended for Hourly Solar Wind Classification

Camporeale et al. (2017) used the combination of hourly averaged solar wind parameters and daily sampled parameters to classify the reference solar wind hourly. However, it is not recommended in this study, at least for the KNN classifier.

We advocate that the time resolution of the reference four-type solar wind and the parameters used for classification should be the same. The temporal resolution of the reference solar wind data sets and the other five solar wind parameters are sampled hourly. However, both *SSN* and *F*10.7 are sampled daily. This means that although the *SSN* and *F*10.7 values are given every hour, their values remain the same for 1 day. Camporeale et al. (2017) used both *SSN* and *F*10.7 to make the classification performance significantly increase but did not explain why leaving out one of the two attributes led to a poorer performance, although both *SSN* and *F*10.7 are known to be strongly correlated, as mentioned by the authors. The hourly repeated daily sampled *SSN* and *F*10.7 parameters seem to be questionable for doing the solar wind classification based on the hourly measured reference solar wind data. Taking only an *SSN*-*F*10.7 two-parameter scheme, for example, the overall accuracy obtained by our KNN classifier is 98.5% and is 99.5%, 98.9%, 98.1%, and 96.6%, for CHOP, SBP, SRRP, and EJECT, respectively. However, this high performance does not indicate that such an *SSN*-*F*10.7 two-parameter scheme is a better choice in solar wind classification.

As shown in Figure 5, it is quite difficult to distinguish CHOP, SBP, SRRP, and EJECT from each other in the plot of *SSN* versus *F*10.7. At the same time, the corresponding decision boundaries for each solar wind category are too complicated to eliminate the concerns of overfitting. One plausible reason is the mismatch of daily sampled parameters and hourly reference solar wind. For the reference solar wind, there are 361 continuous-time segments of CHOP, SBOP, SRRP, and EJECT, with an average time duration of about 24.9 hr. There are 8,625 independent cases when the resolution is set to be 1 hr, while there are only 479 independent cases when the resolution is set to be 1 day. Since the actual resolution of *SSN* and *F*10.7 is 1 day, it might be incorrect to classify the solar wind hourly based on two daily sampled parameters. The ratio of independent data (479/8,625, <6%) is much less than the ratio of training data set (75% in our work and 10–25% in Camporeale et al., 2017). This may result in a lager risk of overfitting. For comparison, the distribution of reference solar wind in the plot of M_f versus S_p is also shown in Figure 5. Although the overall accuracy given by the KNN classifier for M_f-S_p two-parameter scheme is 79.2%, much lower than that for *SSN*-*F*10.7 scheme, it is still possible to generally distinguish the distribution of CHOP, SBP, SRRP, and EJECT from each other, except a few overlaps. The decision boundaries are also more likely to represent a regularized classification.

As shown before, hourly repeated daily sampled *SSN* and *F*10.7 may increase the probability of overfitting. It may be true that the possibility of overfitting gets smaller if more dimensions are considered. Table 5 shows the best performance of solar wind classification including the pair *SSN*-*F*10.7 in different dimensional parameter space. The overall accuracy generally decreases when more parameters are used. However, it is still difficult to quantitatively evaluate the reliable effect of including them on classification performance. Thus, a better choice is not to introduce these two parameters at all, at least for the KNN classifier. Here, we strongly suggest to use the solar wind parameters with the same time resolution as the identified reference solar wind when training the classifier.

4.2. Composition Information in Solar Wind Classification

In the previous classification schemes in two- or three- dimensional parameter space, solar wind composition measurement indeed plays an important role in solar wind classification. However, it is still difficult to

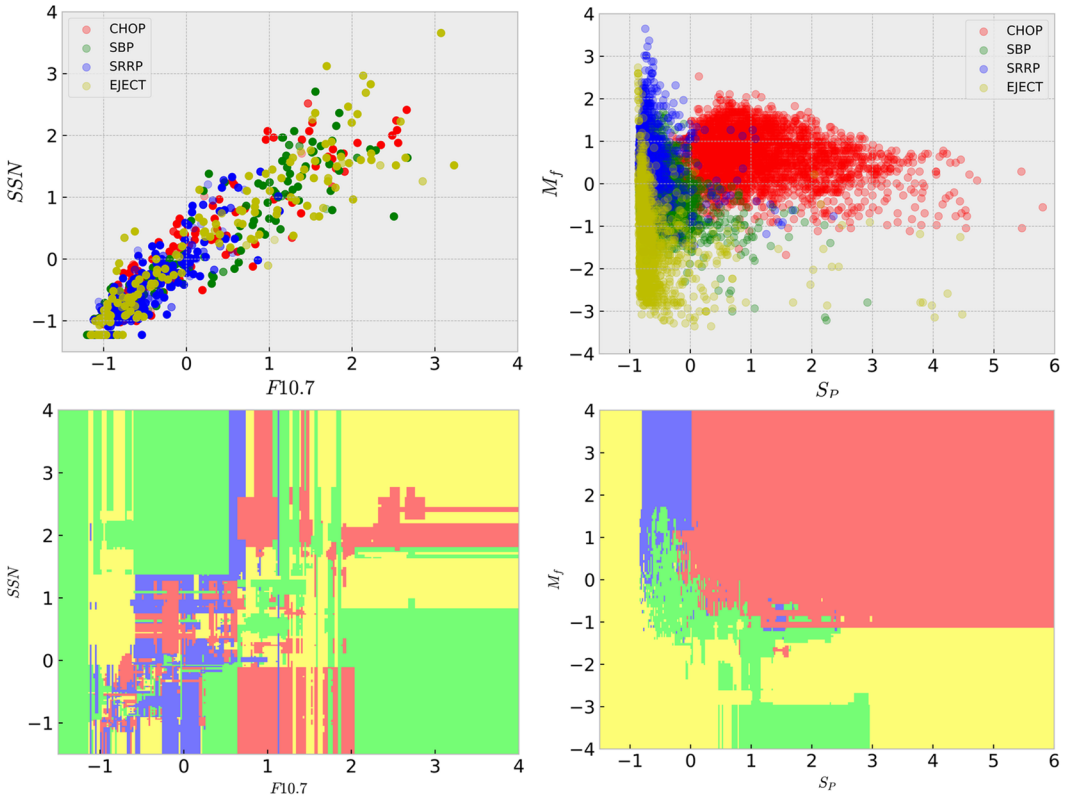

Figure 5. Top: distribution of reference solar wind in the plot of *SSN* versus *F*10.7and M_f versus S_p. Bottom: corresponding decision boundaries for each solar wind category. The overall accuracy given by KNN classifier under the *SSN*-*F*10.7 scheme is 98.5%; however, the accuracy is 79.2% under the M_f-S_p scheme.

conclude that the composition measurement is thus indispensable. To show the importance of composition information in solar wind classification, we have accessed the 1-hr composition data (C^{6+}/C^{5+} and O^{7+}/O^{6+}) from ACE satellite during 1998–2011. During this time interval, the reference solar wind data sets with data gap removed are 8,021 hr: CHOP (2,881 hr), SBOP (2,215 hr), SRRP (1,694 hr), and EJECT (1,231 hr). Compared to the data sets without composition information, the EJECT data reduced from 1,835 to 1,231 hr, and the CHOP, SBOP, SRRP data are the same. The overall classification accuracy by solely using C^{6+}/C^{5+} or O^{7+}/O^{6+} is 51.0% and 65.9%, which is less or comparable to the performance, 66.7%, when S_p is used solely.

The comparison of classification results with/without composition information is shown in Table 6. It is clear that the classification results indeed have some minor improvements, especially when O^{7+}/O^{6+} information is considered. But the improvements are not much significant, only 1.5%. Considering that most of solar wind measuring satellites, for example, the recent Parker Solar Probe, do not have composition instrument, it is suggested that solar wind classification scheme without composition information is still useful.

Table 5
Best Performance of Solar Wind Classification Including the Pair SSN-F10.7 in Different Dimensional Parameter Space

Parameter number	CHOP	SBP	SRRP	EJECT	4-type
0	99.5	98.9	98.1	96.6	98.5
2	99.9	97.3	94.1	93.5	96.7
4	99.8	96.0	93.0	93.5	96.2
6	99.7	95.3	90.3	94.1	95.6
8	99.5	96.4	92.5	94.4	96.2

Note. *SSN* and *F*10.7 have been excluded when counting the parameter number.

Table 6
Comparison of Solar Wind Classification With/Without Composition Information

	CHOP	SBP	SRRP	EJECT	4-type	HKSS
Without composition	99.3	91.4	85.1	92.5	93.1	0.903
C^{6+}/C^{5+}	99.3	92.5	85.6	92.6	93.5	0.909
O^{7+}/O^{6+}	99.4	93.0	89.0	94.1	94.6	0.925
C^{6+}/C^{5+} and O^{7+}/O^{6+}	99.4	93.2	87.3	93.1	94.3	0.920

4.3. Importance of the Accuracy of Reference Solar Wind

The reference solar wind with known types is very important for supervised machine learning. In this study, the reference solar wind data come from expert human knowledge, which may have some uncertainties, especially at the boundaries of events. A natural thought is that the center part of an event has the highest probability to be correctly labeled. For practice, if 3-hr data points at both boundaries were deleted for each EJECT event, the classification accuracy of EJECT should have an improvement of 2.2%. Thus, further improvement of classification accuracy by machine learning is limited by the uncertainties of the reference solar wind data set.

5. Application in Space Weather Early Warning

Solar wind origin information may be helpful for space weather early warning. First, the solar wind category is useful for the risk evaluation of a predicted geomagnetic storm. Turner et al. (2009) showed that the storm intensity and occurrence rate of intense storm (*Dst* minimum < −100 nT) for ICME-driven storms are larger than that for CIR-driven storms. From the storm catalog of Turner et al. (2009), the average *Dst* minimum during a CIR-driven storm is ∼ −75 nT, and the occurrence rate of intense storms is only 13%. However, these two values are ∼ −125 nT and 57% for ICME-driven storms, respectively. Moreover, all superstorms, with *Dst* minimum < −300 nT and midday magnetopause shifting earthward of geosynchronous orbit (Li et al., 2010), are associated with ICMEs. Second, the classification of CHOP and EJECT is also helpful for the risk evaluation of surface charging of geosynchronous spacecraft. Borovsky and Denton (2006) and Denton et al. (2006) found that the magnitude of spacecraft potential is, on average, significantly elevated for CIR-driven storms than during ICME-driven storms. Third, McGranaghan et al. (2014) showed that SBP and SRRP produce forecastable changes in thermospheric density.

Gonzalez and Tsurutani (1987) suggested that storm intensity depends on the intensity of southward interplanetary magnetic field, B_Z, and the threshold for intense storms is summarized to be −10 nT. Echer et al. (2008) later found that storm intensity depends on the solar wind electric field, E_Y, and the threshold for intense storms is summarized to be 5 mV/m. If $B_Z \leq -10$ nT and $E_Y \geq 5$ mV/m are observed in the solar wind at L1 point, a magnetic storm is likely to occur in the next several hours. With the information of solar wind type obtained, more details of the geoeffectiveness can be inferred. Table 7 gives two examples. For the first case, B_Z is observed to be −11.2 nT on 00:00 27 February 2003; moreover, the corresponding E_Y is observed to be 5.03 mV/m. Based on our classification algorithm, the solar wind plasma is categorized to be SBP, indicating a possible CIR-driven storm. Borovsky and Denton (2013) indeed identified that event as a pseudostreamer CIR. Thus, the impending storm will be predicted to likely be a moderate storm with a high risk of dangerous spacecraft surface charging. As a validation, the real occurred storm is identified to be a moderate storm, with the *Dst* minimum of −60 nT. In addition, the magnitude of spacecraft potential (Φ) in geosynchronous orbit during this storm is close to 4,000 V. For the second case, similar B_Z and E_Y are observed to be −10.8 nT and 5.08 mV/m on 00:00 8 November 1998. However, unlikely, the solar wind plasma is categorized to be EJECT for this case (which is later identified as an ICME by Richardson and Cane, https://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm), indicating a possible ICME-driven storm. Thus, the impending storm will likely be an intense storm; however, the risk of spacecraft surface charging is predicted to be relatively low. In fact, the following storm has an intensity of −149 nT, and the magnitude of spacecraft potential during this storm is no more than 900 V.

At present, we use the in situ observation at L1 point to classify the solar wind and can produce a space weather early warning by approximately half an hour. There could be more utility for the present classification scheme if a solar wind monitor is placed at L5. Furthermore, we are still working on improving the prediction window of solar wind classification through utilization of observations on the Sun's surface.

Table 7
Application of the Information of Solar Wind Origin in Improving Space Weather Forecast

Time	B_Z	E_Y	Type	Forecast	$Dst_{\min}$	Φ
27 February 2003	−11.2	5.03	SBP	Moderate CIR-storm	−60	4,000
00:00 UT				high-charging risk		
8 November 1998	−10.8	5.08	EJECT	Intense ICME-storm	−149	900
00:00 UT				low-charging risk		

Note. B_Z denotes the z component of interplanetary magnetic field. E_Y denotes the y component of solar wind electric field. $Dst_{\min}$ denotes the minimum of Dst index during a magnetic storm. Φ denotes the magnitude of spacecraft potential, which is from the Magnetospheric Plasma Analyzer instrument onboard the LANL satellite series.

6. Summary

Solar wind categorization is conducive to understanding the solar wind origin and physical processes ongoing at the Sun. In the face of a great deal of spacecraft observations, manual classification by domain knowledge experts is prohibitive in terms of time and subject to human error. Thus, automatic classification methods are needed. Recently, with the rapid developments in the field of artificial intelligence, classification by machine learning is becoming more and more popular and powerful in big-volume data analysis, and furthermore, its performance is improving as well.

In this study, 10 additional popular supervised machine learning models, KNN, LSVM, RBFSVM, DT, RF, AdaBoost, NN, GNB, QDA, and XGBoost, are used to classify the solar wind at 1 AU into four plasma types: CHOP, SBP, SRRP, and EJECT.

A total of 13 parameters, each with 1-hr temporal resolution, are used for training the classifiers and searching for the best variable scheme. These parameters are the magnetic field intensity B_T, the proton number density N_P, the proton temperature T_P, the solar wind speed V_P, the proton-specific entropy S_p, the Alfvén speed V_A, the ratio of velocity-dependent expected proton temperature and proton temperature $T_{\exp}/T_P$, the number density ratio of proton and alpha $N_{\alpha p}$, the dynamic pressure P_d, the solar wind electric field E_y, the plasma beta value β, the Alfvén Mach number M_A, and the fast magnetosonic Mach number M_f. Note that all the parameters can be obtained or derived from the typical solar wind observations. No composition measurements are needed, allowing our algorithm to be applied to most solar wind measuring spacecraft.

By exhaustive enumeration, an eight-dimensional scheme (B_T, N_P, T_P, V_P, $N_{\alpha p}$, $T_{\exp}/T_P$, S_p, and M_f) is found to obtain the highest classification accuracy among all the 8,191 combinations of the above 13 parameters. Among the 10 popular classifiers, the KNN classifier obtains the best accuracy of 92.8%. It significantly improves the accuracy over existing schemes that only use the in situ solar wind magnetic field and plasma observations, as done by Zhao et al. (2009) and Xu and Borovsky (2015). Although the accuracy obtained by our KNN classifier method is 5.4% less than that of Camporeale et al. (2017), their mixture of hourly averaged solar wind parameters and daily sampled parameters is not recommended here because of an enhanced risk of overfitting as discussed in section 4.1. Other machine learning classifiers, such as XGBoost, RF, RBFSVM, and NN, also perform well in solar wind classification, with the accuracy greater than 91.0%. These results can enhance people's confidence in using machine learning techniques for solar wind classification.

Although the accuracy can be improved by 1.5% when O^{7+}/O^{6+} information is additionally considered, the scheme presented here without composition is still good enough and could be applicable for solar wind measuring spacecraft. Small-scale flux rope events may also be identifiable based on our method, though further investigation and validation are needed. In addition, two application examples of solar wind classification are given, indicating that it may be helpful for the risk evaluation of predicted magnetic storms and surface charging of geosynchronous spacecraft.

This work emphasizes the classification technique itself rather than the science of the solar wind origin. More efforts by the community are needed to bring about further understanding in the science aspects. In the future, with potentially new solar wind types and corresponding new event data, our machine learning approach will be updated, accordingly.

Acknowledgments
We thank the OMNI database for the use of solar wind data, which is accessible in the website (https://spdf.gsfc.nasa.gov/pub/data/omni/low_res_omni/). We also thank the scikit-learn and XGBoost toolkits written in Python, which provide the classification classifiers used here and can be found in the website (https://scikit-learn.org/stable/install.html). The reference solar wind data for training and testing used in this study and the final classified solar wind can be accessed in the website (https://www.spaceweather.ac.cn/%7Ehli/research.html). This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (CAS) (Grants XDA17010301 and XDA15052500), National Natural Science Foundation of China (NNSFC) (Grants 41874203, 41574169, 41574159, and 41731070), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) (2016QNRC001), and Key Research Program of Frontier Sciences, CAS (QYZDJ-SSW-JSC028). H. Li was also supported by the project of Civil Aerospace "13th Five Year Plan" Preliminary Research in Space Science (project name: Research on Important Scientific Issues of Heliospheric Boundary Exploration, Project No. D020301), Youth Innovation Promotion Association of the Chinese Academy of Sciences, and in part by the Specialized Research Fund for State Key Laboratories of China.

References

Antiochos, S. K., Mikić, Z., Titov, V. S., Lionello, R., & Linker, J. A. (2011). A model for the sources of the slow solar wind. *The Astrophysical Journal*, *731*(2), 112. https://doi.org/10.1088/0004-637X/731/2/112

Antonucci, E., Abbo, L., & Dodero, M. A. (2005). Slow wind and magnetic topology in the solar minimum corona in 1996–1997. *Astronomy and Astrophysics*, *435*(2), 699–711. https://doi.org/10.1051/0004-6361:20047126

Arge, C. N., Odstrcil, D., Pizzo, V. J., & Mayer, L. R. (2003). Improved method for specifying solar wind speed near the Sun. Paper presented at the Solar Wind Ten, https://doi.org/10.1063/1.1618574

Arya, S., & Freeman, J. W. (2012). Estimates of solar wind velocity gradients between 0.3 and 1 AU based on velocity probability distributions from Helios 1 at perihelion and aphelion. *Journal of Geophysical Research*, *96*(A8), 14,183–14,187. https://doi.org/10.1029/91JA01135

Asbridge, J. R., Bame, S. J., Feldman, W. C., & Montgomery, M. D. (1976). Helium and hydrogen velocity differences in the solar wind. *Journal of Geophysical Research*, *81*(16), 2719–2727. https://doi.org/10.1029/JA081i016p02719

Bame, S. J., Asbridge, J. R., Feldman, W. C., & Gosling, J. T. (1977). Evidence for a structure-free state at high solar wind speeds. *Journal of Geophysical Research*, *82*, 1487–1492. https://doi.org/10.1029/JA082i010p01487

Borovsky, J. E. (2008). Flux tube texture of the solar wind: Strands of the magnetic carpet at 1 AU? *Journal of Geophysical Research*, *113*, A08110. https://doi.org/10.1029/2007JA012684

Borovsky, J. E. (2010). On the variations of the solar wind magnetic field about the Parker spiral direction. *Journal of Geophysical Research*, *115*, A09101. https://doi.org/10.1029/2009JA015040

Borovsky, J. E. (2012). The velocity and magnetic field fluctuations of the solar wind at 1 AU: Statistical analysis of Fourier spectra and correlations with plasma properties. *Journal of Geophysical Research*, *117*, A05104. https://doi.org/10.1029/2011JA017499

Borovsky, J. E., & Denton, M. H. (2006). Differences between CME-driven storms and CIR-driven storms. *Journal of Geophysical Research*, *111*, A07S08. https://doi.org/10.1029/2005JA011447

Borovsky, J. E., & Denton, M. H. (2013). The differences between storms driven by helmet streamer CIRs and storms driven by pseudostreamer CIRs. *Journal of Geophysical Research: Space Physics*, *118*, 5506–5521. https://doi.org/10.1002/jgra.50524

Borovsky, J. E., & Denton, M. H. (2014). Exploring the cross correlations and autocorrelations of the ULF indices and incorporating the ULF indices into the systems science of the solar wind-driven magnetosphere. *Journal of Geophysical Research: Space Physics*, *119*, 4307–4334. https://doi.org/10.1002/2014JA019876

Bothmer, V., & Schwenn, R. (1996). Signatures of fast CMEs in interplanetary space. *Advances in Space Research*, *17*(4), 319–322. https://doi.org/10.1016/0273-1177(95)00593-4

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software.

Buhmann, M. D. (2003). *Radial basis functions: Theory and Implementations*. Cambridge: Cambridge University Press.

Cady, F. (2017). Machine learning classification, *The data science handbook*. Hoboken, New Jersey: John Wiley and Sons Inc. https://doi.org/10.1002/9781119092919.ch8

Camporeale, E., Carè, A., & Borovsky, J. E. (2017). Classification of solar wind with machine learning. *Journal of Geophysical Research: Space Physics*, *122*, 10,910–10,920. https://doi.org/10.1002/2017JA024383

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Crooker, N. U., Antiochos, S. K., Zhao, X., & Neugebauer, M. (2012). Global network of slow solar wind. *Journal of Geophysical Research*, *117*, A04104. https://doi.org/10.1029/2011JA017236

Denoeux, T. (1995). A *k*-nearest neighbor classification rule-based on Dempster-Shafer theory. *IEEE Transactions on Systems Man and Cybernetics*, *25*(5), 804–813. https://doi.org/10.1109/21.376493

Denton, M. H., Borovsky, J. E., Skoug, R. M., Thomsen, M. F., Lavraud, B., Henderson, M. G., et al. (2006). Geomagnetic storms driven by ICME- and CIR-dominated solar wind. *Journal of Geophysical Research*, *111*, A07S07. https://doi.org/10.1029/2005JA011436

Echer, E., Gonzalez, W. D., Tsurutani, B. T., & Gonzalez, A. L. C. (2008). Interplanetary conditions causing intense geomagnetic storms ($Dst \leq -100$ nT) during solar cycle 23 (1996–2006). *Journal of Geophysical Research*, *113*, A05221. https://doi.org/10.1029/2007JA012744

Eyni, M., & Steinitz, R. (1978). Cooling of slow solar wind protons from the Helios 1 experiment. *Journal of Geophysical Research*, *83*, 4387. https://doi.org/10.1029/JA083iA09p04387

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research.*, *9*, 1871–1874.

Feldman, U., Landi, E., & Schwadron, N. A. (2005). On the sources of fast and slow solar wind. *Journal of Geophysical Research*, *110*, A07109. https://doi.org/10.1029/2004JA010918

Fisk, L. A., Zurbuchen, T. H., & Schwadron, N. A. (1999). On the coronal magnetic field: Consequences of large-scale motions. *The Astrophysical Journal*, *521*, 868–877. https://doi.org/10.1086/307556

Foullon, C., Lavraud, B., Luhmann, J. G., Farrugia, C. J., Retinò, A., Simunac, K. D. C., et al. (2011). Plasmoid releases in the heliospheric current sheet and associated coronal hole boundary layer evolution. *The Astrophysical Journal*, *737*(1), 16. https://doi.org/10.1088/0004-637X/737/1/16

Gonzalez, W. D., & Tsurutani, B. T. (1987). Criteria of interplanetary parameters causing intense magnetic storms ($D_{st} < -100$ nT. *Planetary and Space Science*, *35*, 1101.

Gosling, J. T., Borrini, G., Asbridge, J. R., Bame, S. J., Feldman, W. C., & Hansen, R. T. (2012). Coronal streamers in the solar wind at 1 AU. *Journal of Geophysical Research*, *86*, 5438–5448. https://doi.org/10.1029/JA086iA07p05438

Hellinger, P., Matteini, L., Stverák, S., Trávnícek, P. M., & Marsch, E. (2011). Heating and cooling of protons in the fast solar wind between 0.3 and 1 AU: Helios revisited. *Journal of Geophysical Research*, *116*, A09105. https://doi.org/10.1029/2011JA016674

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14-16 August 1995, pp. 278–282.

Jian, L., Russell, C. T., Luhmann, J. G., & Skoug, R. M. (2006). Properties of interplanetary coronal mass ejections at one AU during 1995–2004. *Solar Physics*, *239*(1-2), 393–436. https://doi.org/10.1007/s11207-006-0133-2

Klein, L. W., & Burlaga, L. F. (1982). Interplanetary magnetic clouds at 1 AU. *Journal of Geophysical Research*, *87*, 613–624. https://doi.org/10.1029/JA087iA02p00613

Kunow, H., Crooker, N. U., Linker, J. A., Schwenn, R., & Von Steiger, R. (2006). *Coronal mass ejections*, pp. 484. Dordrechtl; Norwell, MA: Springer.

Lepping, R. P., Wu, C.-C., & Berdichevsky, D. B. (2005). Automatic identification of magnetic clouds and cloud-like regions at 1 AU: Occurrence rate and other properties. *Annales Geophysicae*, *23*(7), 2687–2704. https://doi.org/10.5194/angeo-23-2687-2005

Li, G., Miao, B., Hu, Q., & Qin, G. (2011). Effect of current sheets on the solar wind magnetic field power spectrum from the Ulysses observation: From Kraichnan to Kolmogorov scaling. *Physical Review Letters*, *106*(12), 125,001. https://doi.org/10.1103/PhysRevLett.106.125001

Li, H., Wang, C., He, J., Zhang, L., Richardson, J. D., Belcher, J. W., & Tu, C. (2016). Plasma heating inside interplanetary coronal mass ejections by Alfvénic fluctuations dissipation. *The Astrophysical Journal Letters*, *831*(2), L13. https://doi.org/10.3847/2041-8205/831/2/L13

Li, H., Wang, C., & Kan, J. R. (2010). Midday magnetopause shifts earthward of geosynchronous orbit during geomagnetic superstorms with Dst ≤ -300 nT. *Journal of Geophysical Research*, *115*, A08230. https://doi.org/10.1029/2009JA014612

Li, H., Wang, C., Richardson, J. D., & Tu, C. (2017). Evolution of Alfvénic fluctuations inside an interplanetary coronal mass ejection and their contribution to local plasma heating: Joint observations from 1.0 to 5.4 au. *The Astrophysical Journal*, *851*(1), L2. https://doi.org/10.3847/2041-8213/aa9c3f

Luttrell, A. H., & Richter, A. K. (1988). The role of Alfvénic fluctuations in MHD turbulence evolution between 0.3 and 1 AU. In V. J. Pizzo, T. E. Holzer, & D. G. Sime (Eds.), *Proceedings of the Sixth International Solar Wind Conference* (pp. 335). Boulder, Colo.

Mariani, F., Bavassano, B., & Villante, U. (1983). A statistical study of MHD discontinuities in the inner solar system: Helios 1 and 2. *Solar Physics*, *83*, 349–365. https://doi.org/10.1007/BF00148285

Marsch, E., Rosenbauer, H., Schwenn, R., Muehlhaeuser, K.-H., & Neubauer, F. M. (1982). Solar wind helium ions-observations of the Helios solar probes between 0.3 and 1 AU. *Journal of Geophysical Research*, *87*, 35–51. https://doi.org/10.1029/JA087iA01p00035

Matthaeus, W. H., Breech, B., Dmitruk, P., Bemporad, A., Poletto, G., Velli, M., & Romoli, M. (2007). Density and magnetic field signatures of interplanetary $1/f$ noise. *The Astrophysical Journal Letters*, *657*(2), L121. https://doi.org/10.1086/513075

McComas, D. J., Ebert, R. W., Elliott, H. A., Goldstein, B. E., Gosling, J. T., Schwadron, N. A., & Skoug, R. M. (2008). Weaker solar wind from the polar coronal holes and the whole sun. *Geophysical Research Letters*, *35*, L18103. https://doi.org/10.1029/2008GL034896

McGranaghan, R., Knipp, D. J., McPherron, R. L., & Hunt, L. A. (2014). Impact of equinoctial high-speed stream structures on thermospheric responses. *Space Weather*, *12*, 277–297. https://doi.org/10.1002/2014SW001045

Moldwin, M. B., Ford, S., Lepping, R., Slavin, J., & Szabo, A. (2000). Small-scale magnetic flux ropes in the solar wind. *Geophysical Research Letters*, *27*(1), 57–60. https://doi.org/10.1029/1999GL010724

Neugebauer, M., Steinberg, J. T., Tokar, R. L., Barraclough, B. L., Dors, E. E., & Wiens, R. C. (2003). Genesis on-board determination of the solar wind flow regime. In C. T. Russell (Ed.), *The Genesis mission* (pp. 153–171). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0241-7_6

Newbury, J. A., Russell, C. T., Phillips, J. L., & Gary, S. P. (1998). Electron temperature in the ambient solar wind: Typical properties and a lower bound at 1 AU. *Journal of Geophysical Research*, *103*(A5), 9553–9566. https://doi.org/10.1029/98JA00067

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, & Varoquaux, G. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.3389/fninf.2014.00014

Perez, A., Larranaga, P., & Inza, I. (2006). Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, *43*(1), 1–25.

Reisenfeld, D. B., Steinberg, J. T., Barraclough, B. L., Dors, E. E., Wiens, R. C., Neugebauer, M., et al. (2003). Comparison of the Genesis solar wind regime algorithm results with solar wind composition observed by ACE. *AIP Conference Proceedings*, *679*(1), 632–635. https://doi.org/10.1063/1.1618674

Richardson, I. G., & Cane, H. V. (2004). The fraction of interplanetary coronal mass ejections that are magnetic clouds: Evidence for a solar cycle variation. *Geophysical Research Letters*, *31*, L18804. https://doi.org/10.1029/2004GL020958

Richardson, I. G., & Cane, H. V. (2010). Near-earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009) catalog and summary of properties. *Solar Physics*, *264*(1), 189–237. https://doi.org/10.1007/s11207-010-9568-6

Richardson, I. G., Cliver, E. W., & Cane, H. V. (2000). Sources of geomagnetic activity over the solar cycle: Relative importance of coronal mass ejections, high-speed streams, and slow solar wind. *Journal of Geophysical Research*, *105*, 18,203–18,213. https://doi.org/10.1029/1999JA000400

Rojas, R. (1996). *Neural networks: A systematic introduction*. Berlin, New-York: Springer-Verlag.

Schwenn, R. (1990). Large scale structure of the interplanetary medium. In R. Schwenn & E. Marsch (Eds.), *Physics of the inner heliosphere I* pp. 99). Berlin: Springer.

Schwenn, R. (2006). Solar wind sources and their variations over the solar cycle. *Space Science Reviews*, *124*(1-4), 51–76. https://doi.org/10.1007/s11214-006-9099-5

Sheeley, N. R., Harvey, J. W., & Feldman, W. C. (1976). Coronal holes, solar wind streams, and recurrent geomagnetic disturbances: 1973–1976. *Solar Physics*, *49*(2), 271–278. https://doi.org/10.1007/BF00162451

Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, *8*, 1277–1305.

Subramanian, S., Madjarska, M. S., & Doyle, J. G. (2010). Coronal hole boundaries evolution at small scales: II. XRT view. Can small-scale outflows at CHBs be a source of the slow solar wind? *Astronomy and Astrophysics*, *516*, A50. https://doi.org/10.1051/0004-6361/200913624

Suess, S. T., Ko, Y.-K., von Steiger, R., & Moore, R. L. (2009). Quiescent current sheets in the solar wind and origins of slow wind. *Journal of Geophysical Research*, *114*, A04103. https://doi.org/10.1029/2008JA013704

Thieme, K. M., Marsch, E., & Schwenn, R. (1990). Spatial structures in high-speed streams as signatures of fine structures in coronal holes. *Annales Geophysicae*, *8*, 713–723.

Thieme, K. M., Schwenn, R., & Marsch, E. (1989). Are structures in high-speed streams signatures of coronal fine structures? *Advances in Space Research*, *9*, 127–130. https://doi.org/10.1016/0273-1177(89)90105-1

Tu, C.-Y., & Marsch, E. (1995). MHD structures, waves and turbulence in the solar wind: Observations and theories. *Space Science Reviews*, *73*(1-2), 1–210.

Turner, N. E., Cramer, W. D., Earles, S. K., & Emery, B. A. (2009). Geoefficiency and energy partitioning in CIR-driven and CME-driven storms. *Journal of Atmospheric and Solar-Terrestrial Physics*, *71*(10-11), 1023–1031. https://doi.org/10.1016/j.jastp.2009.02.005

von Steiger, R., Zurbuchen, T. H., & McComas, D. J. (2010). Oxygen flux in the solar wind: Ulysses observations. *Geophysical Research Letters*, *37*, L22101. https://doi.org/10.1029/2010GL045389

Wang, Y.-M., & Sheeley, N. R. (1990). Solar wind speed and coronal flux-tube expansion. *The Astrophysical Journal*, *355*, 726–732. https://doi.org/10.1086/168805

Xu, F., & Borovsky, J. E. (2015). A new four-plasma categorization scheme for the solar wind: 4-plasma solar-wind categorization. *Journal of Geophysical Research: Space Physics*, *120*, 70–100. https://doi.org/10.1002/2014JA020412

Yordanova, E., Balogh, A., Noullez, A., & von Steiger, R. (2009). Turbulence and intermittency in the heliospheric magnetic field in fast and slow solar wind: Turbulence and intermittency in the solar wind. *Journal of Geophysical Research*, *114*, A08101. https://doi.org/10.1029/2009JA014067

Zastenker, G. N., Koloskova, I. V., Riazantseva, M. O., Yurasov, A. S., Safrankova, J., Nemecek, Z., et al. (2014). Observation of fast variations of the helium-ion abundance in the solar wind. *Cosmic Research*, *52*(1), 25–36. https://doi.org/10.1134/S0010952514010109

Zhao, L., Zurbuchen, T. H., & Fisk, L. A. (2009). Global distribution of the solar wind during solar cycle 23: ACE observations. *Geophysical Research Letters*, *36*, L14104. https://doi.org/10.1029/2009GL039181

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class adaboost. *Statistics and Its Interface*, *2*, 349–360.

Zurbuchen, T. H., Fisk, L. A., Gloeckler, G., & von Steiger, R. (2002). The solar wind composition throughout the solar cycle: A continuum of dynamic states. *Geophysical Research Letters*, *29*, 1352. https://doi.org/10.1029/2001GL013946

Zurbuchen, T. H., & Richardson, I. G. (2006). In-situ solar wind and magnetic field signatures of interplanetary coronal mass ejections. *Space Science Reviews*, *123*(1-3), 31–43. https://doi.org/10.1007/s11214-006-9010-4