RAA 2017 Vol. **17** No. **11**, 116 (14pp) doi: 10.1088/1674–4527/17/11/116

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Short-term solar flare prediction using image-case-based reasoning

Jin-Fu Liu[1], Fei Li[2], Huai-Peng Zhang[1] and Da-Ren Yu[1]

[1] School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China
[2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; *feili_hit@yahoo.com*

Received 2017 April 12; accepted 2017 August 3

Abstract Solar flares strongly influence space weather and human activities, and their prediction is highly complex. The existing solutions such as data based approaches and model based approaches have a common shortcoming which is the lack of human engagement in the forecasting process. An image-case-based reasoning method is introduced to achieve this goal. The image case library is composed of *SOHO*/MDI longitudinal magnetograms, the images from which exhibit the maximum horizontal gradient, the length of the neutral line and the number of singular points that are extracted for retrieving similar image cases. Genetic optimization algorithms are employed for optimizing the weight assignment for image features and the number of similar image cases retrieved. Similar image cases and prediction results derived by majority voting for these similar image cases are output and shown to the forecaster in order to integrate his/her experience with the final prediction results. Experimental results demonstrate that the case-based reasoning approach has slightly better performance than other methods, and is more efficient with forecasts improved by humans.

Key words: methods: statistical — Sun: activity — Sun: magnetic fields — Sun: photosphere — Sun: flares

1 INTRODUCTION

Predicting the eruption of solar flares is practically significant due to their potential threat. Many efforts have been devoted into improving the performance of short-term flare predictions. Hitherto, expert systems, statistical models and artificial intelligence techniques have been employed for solar flare prediction.

One aspect of this work is to find more informative predictors. McIntosh (1990) proposed the McIntosh classification of sunspots to reflect the morphological characteristics of active regions (ARs) and developed an expert system called Theophrastus to predict solar flares. This classification scheme and system have been widely applied in subsequent research. Bornmann & Shaw (1994) used multiple linear regression analysis to derive the effective contributions to solar flare prediction for each parameter and they concluded that when reduced to 10 parameters, the observed flare rates can still be adequately replicated. McIntosh (1990) discovered that the first three of these accurately represented the length of the sunspot group, the size and shape of the largest spot, and the distribution of spots within that group. The accumulation of historical flare rates from the McIntosh classification system led to ideas for estimating flare probability. Based on the assumption that solar flare eruptions obey a Poisson distribution in time and a power-law distribution in size, a statistical prediction method was formulated by Wheatland (2001). Gallagher et al. (2002) evaluated the prediction rate using the McIntosh classification system, and then forecast the occurrence of daily flares by assuming a Poisson distribution for the waiting time of X-ray flares. Bloomfield et al. (2012) used X-ray flares measured by the *Geostationary Operational Environment Satellite* and McIntosh group classifications to determine the Poisson probabilities for different flare magnitudes. Cui et al. (2006) presented three types of predictors to describe the nonpotentiality and complexity

of a photospheric magnetic field: the maximum horizontal gradient, the length of the neutral line and the number of singular points. Yu et al. (2010a) implemented a model that applies multiple-resolution predictors resulting from the decomposition of a sequence of predictors into four frequency bands using a maximal overlap discrete wavelet transform. This model reflects the trend and the changing rate of emerging flux regions. Huang et al. (2010) constructed a C4.5 decision tree model based on predictor teams that are extracted from a dataset using rough set theory. The predictor teams used in the ensemble model not only efficiently reduce redundancy but also increase profitability of the relevant information. Huang et al. (2013) constructed a metric to depict the positional relationship between ARs and predicted active longitudes, enhancing the performance of solar flare prediction. Volobuev et al. (2016) proposed that a generalized Laplacian could help predict strong flares and found that the maximum Laplacian is located near the AR polarity inversion line.

Another aspect of solar flare prediction research is to establish more powerful models. Wheatland (2004) applied a Bayesian approach capable of refining the initial prediction using the prior probability of the prediction, the flaring records and the phenomenological rules of flare statistics. Leka & Barnes (2003a) and Leka & Barnes (2003b) employed Fisher's linear discriminant analysis to predict whether a flare will occur. Li et al. (2007) constructed a prediction model based on the combination of a support vector machine (SVM) and a k-nearest neighbors (KNN) algorithm. Combining an SVM and a cascade correlation neural network, Qahwaji & Colak (2007) put forward a hybrid system for automatic detection. Wang et al. (2008) proposed new measurements based on solar magnetic field observations that provide more information than what is given by measurements based on the sunspot group classification and next they set up a solar flare forecasting model supported by an artificial neural network. To predict short-term significant solar flares, Colak & Qahwaji (2009) built an automated hybrid computer platform (ASAP) using *Solar and Heliospheric Observatory* (*SOHO*)/Michelson Doppler Imager (MDI) images. Yu et al. (2010b) presented a Bayesian network approach for short-term solar flare level prediction by extracting sequential features and analyzing their temporal variations with respect to flare eruptions. Li & Sun (2013) considered the evolution of solar ARs and used sequential sunspot data to predict solar flares. Ahmed et al. (2013) applied machine-learning and feature-selection algorithms to a set of magnetic feature properties to determine solar flare prediction capabilities and the relationship between these properties and flare occurrence. Muranushi et al. (2015) developed the UFCORIN platform for studying and automating the prediction of space weather, including solar flares. Boucheron et al. (2015) developed an SVM based approach to predict flare size and occurrence time. Shin et al. (2016) focused on flare flux for strong flares, and proposed daily maximum flare-flux forecast models for strong flares (M- and X-class) using multiple linear regression and artificial neural network methods.

Currently, solar flare prediction mainly concentrates on building powerful prediction models and exploring more informative predictors. Hence Barnes et al. (2016) make a comparison with a number of existing algorithms applied to common data sets, specifically line-of-sight magnetic field and continuum intensity images from MDI. However, prediction results provided by models such as neural networks, SVMs, C4.5 decision trees and Bayesian networks are difficult for a forecaster to understand in spite of the high potential accuracy of their prediction. A common weakness in these models is that they provide the forecaster with little comprehensible information apart from the final prediction results. Case-based reasoning (CBR) is a problem-solving paradigm that solves new problems by analyzing and adapting solutions used for similar past problems (Riesbeck & Schank 2013). CBR simulates human problem-solving methodology through memory activation and inferences, and its primary process is to identify the current case, find previous cases similar to the current one, suggest a solution based on the retrieved similar cases, evaluate the proposed solution and finally to update the system by learning from the current experience.

Figure 1 illustrates the basic flow chart for a CBR system. Nowadays, the application of CBR is popular in health sciences (Begum et al. 2011), academia (Mamaghani 2002), design (Guo et al. 2013), industry (Mikos et al. 2010) and business (Li & Sun 2013) owing its intuitive nature, minimum knowledge requirements, ease of understanding and explanation, and high level of interpretability. While other machine learning techniques generalize associations between features and outcomes, CBR justifies the solution to a new problem through experience accumulated from concrete past situations (Marir & Watson 1994, Richter & Aamodt 2005).

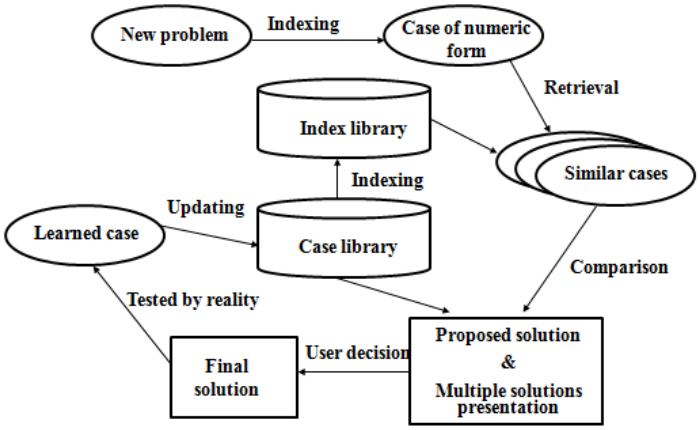

Fig. 1 Basic CBR flow chart. A new problem is generally encoded into numeric form and retrieved by comparison with similar cases. These cases are stored in the case library, which provides relevant solutions and their index. By comparing with similar cases, the CBR system may give some possible solutions for a new problem to help humans to make a final decision which is then tested by comparing with reality. After that, the new problem becomes a new case stored in the case library and is assigned a new index, and can act as a solution for other problems that emerge.

Original observational data used for solar flare prediction take the form of images which provide the most abundant information about flare eruption. CBR in solar flare image cases leads to a more interpretable approach, when compared to models built from numeric data. Therefore, in this article, *SOHO*/MDI longitudinal magnetograms are used to construct the image case library and the image-case-based reasoning method is proposed to predict solar flare eruptions.

The structure of this article is organized as follows. The data are introduced in Section 2. The image case library construction based on *SOHO*/MDI images is presented in Section 3. The genetically optimized similar case retrieval method is proposed in Section 4. Case adaption involving the participation of forecasters and incremental learning is explained in Section 5. Experimental results and analysis of the proposed method are reported in Section 6. Finally, conclusions and future research outlook are provided in Section 7.

2 DATA

SOHO/MDI full disk longitudinal magnetograms were downloaded from *ftp://soi-ftp.stanford.edu/pub /magnetograms/*. The magnetograms used in this study have a spatial resolution of $2'' \times 2''$ after a smoothing average of 3×4 pixels for data reduction. The noise level of the line-of-sight magnetic field is less than 20 G (Wang et al. 1996). These images were collected from 1996 April 15 to 2008 April 2. The daily magnetograms were recorded at intervals of 96 minutes, giving 15 synoptic maps each day. In FITS format, the *SOHO*/MDI images were recorded on an array of 1024×1024 pixels, and the spatial resolution of a *SOHO*/MDI magnetogram is $4''$ over the whole solar disk. Figure 2 shows a *SOHO*/MDI magnetogram (2003 November 23).

AR location data associated with the solar flare events are obtained from *ftp://ftp.swpc.noaa.gov/pub/warehouse/*. We only consider the region within 30° of the solar disk center for our prediction and data analysis. Moreover, all the data in this study are extracted from ARs selected based on the following two criteria:

(1) There exists at least one X-ray flare whose magnitude $\geq$ C1.0.
(2) The location of ARs is within 30°of the solar disk center, where projection effects can be neglected.

The importance of a solar flare is conventionally described by its index, for example, C, M or X. Within the forecasting period, more than one flare may happen. The importance of these flares is summed up with weights. The total importance of flares is computed as follows (Abramenko 2005)

$$I_{\rm tot} = 1 \times \sum C + 10 \times \sum M + 100 \times \sum X. \quad (1)$$

Solar flare samples are extracted based on an AR containing the predictors and the total importance of a

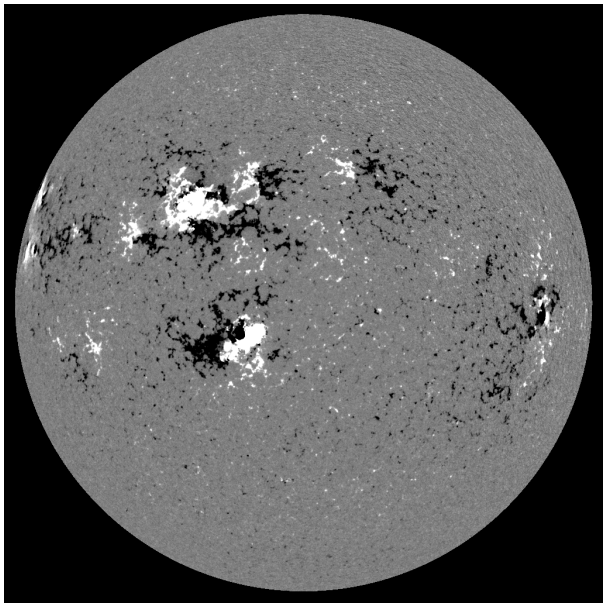

Fig. 2 Full disk magnetogram taken on 2003 November 23 (*SOHO*/MDI data). The *black* and *white* areas represent negative and positive magnetic poles respectively.

solar flare. Total importance ($I_{\rm tot}$) of a sample is calculated using Equation (1). Equation (1) considers the influence of all flares within the forward-looking period. To better illustrate this: if, for example, an AR produces C1.2, C2.3, M4.1 and X1.2 flares within 48 h, $I_{\rm tot} = (1.2{+}2.3){+}10{\times}4.1{+}100{\times}1.2 = 164.5$ (Wang & Japkowicz 2010). Flares with a significance above a certain threshold are often used in forecasting models and the threshold of $I_{\rm tot}$ is set to be 10. Therefore, a flaring sample is defined as one that has a total importance greater than 10.

3 IMAGE CASE LIBRARY CONSTRUCTION USING *SOHO*/MDI MAGNETOGRAMS

A case library, where past cases and their solutions are stored, can be constructed. It is an experience-rich database that contains abundant information. A case can take any form, including a signal, image, video or written document, as long as it can comprise a clear and understandable description of the problems to be addressed. Establishing an informative case library is essential when constructing an effective CBR system. In the field of medical diagnosis, images are widely used for the case library. Liu et al. (2012) built a computer-aided breast cancer diagnosis system using CBR with color Doppler flow images as the case library. Zhou et al. (2012) used fracture images as the case base to assist surgeons in decisions regarding new cases by supplying visually similar past cases. One advantage of using images as the case library that should not be overlooked is their visual presentation, which can support the suggested solution with a more intuitive interpretation. Moreover, images contain the most abundant original information, whereas information provided by numeric predictors is more limited.

Therefore, we use images as the case library for solar flare prediction. In CBR, all predictors are extracted from an AR in full disk images. The information about the AR comes from supplementary materials associated with the cases. The case library consists of *SOHO*/MDI full disk longitudinal magnetograms. Each image and its solution make up a raw case. Images are used to construct the case base, because when a forecaster attempts to adapt solutions to similar cases, magnetograms can provide a direct visualization reference that the forecaster is familiar with. In contrast, numeric predictors do not provide visual help for reference. Combining the suggested solution generated by CBR with the forecaster's experience about similar image cases is helpful for improving prediction accuracy.

Though the magnetograms contain large quantities of information, it is difficult to compare images directly to rank similarities. Hence, image numeric features should be extracted from them to compute the similarity between image cases.

Table 1 Values of Parameters in the Sigmoid Function

Threshold	Forward-looking period	Predictor	A_1	A_2	X_0	W
$I_{\rm tot} = 10$	Within 48 (h)	$\lvert\nabla_h B_z\rvert_m$	0.164	0.738	0.360	0.066
		L	0.062	0.848	763.08	382.97
		η	–0.196	0.730	9.343	22.663

In this paper, the maximum horizontal gradient ($|\nabla_h B_z|_m$), the length of neutral line (L) and the number of singular points (η) are extracted from *SOHO*/MDI magnetograms and then the image numeric features are mapped by a sigmoid function, analogous to a Boltzmann sigmoid, as shown in Equation (2) to incorporate known relations between the features and the flare level (Cui et al. 2006). Other image numeric features can also be chosen to compute the similarity between image cases.

$$Y = A_2 + \frac{A_1 - A_2}{\exp[(X - X_0)/W]}, \tag{2}$$

where Y is the flare productivity defined by the ratio of the number of flaring samples to the total number of samples, and X is one of the image numeric feature values (the length of a neutral line, the number of singular points and the maximum horizontal gradient) for the ARs. A_1, A_2, X_0 and W are optimized in the curve-fitting process to minimize the sum of squares of the deviations between the observed data and the expected data (Ledvij 2003). Table 1 shows optimized values of the parameters.

4 GENETICALLY-OPTIMIZED CASE RETRIEVAL

4.1 Similarity Measurement Using Weighted Euclidean Distance

Osborne & Bridge (1996) suggested some guidelines on theoretical frameworks for systematically constructing similarity measures in CBR. Reliable matching and ranking in the process of comparing two cases largely depend on identifying a suitable similarity measurement. If the metric cannot sufficiently and appropriately differentiate cases, the CBR system yields poor prediction accuracy. It is evident from the current literature on similarity measurements that each metric has its corresponding strengths and weaknesses, and that identifying the most appropriate one depends on the type of problem. Although there is an abundance of metrics, one point to bear in mind is that, for a similarity measurement to be effective and credible, it must take into account the relative importance of features. The more important features contribute more to the aggregation of differences between cases, while the less important ones contribute less. Proper weight assignment can enhance the performance of CBR and decrease its sensitivity to similarity measurement. This means that selecting a suitable similarity measurement with feature weights is an essential element in CBR.

The distance between two cases is the most obvious measure for similarity. Euclidean distance functions are the most commonly-used distance measures. For two cases, A and B, with n numeric feature values, the distance between A and B is given by

$$\text{Euclidean Distance} : \text{DIS}_{AB} = \sqrt{\sum_{k=1}^{n} |A_k - B_k|^2}, \tag{3}$$

where A_k and B_k are the $k^{\rm th}$ numeric feature values of A and B, respectively.

Although Euclidean functions are simple and easy to compute, Berry & Linoff (1997) pointed out that they have desirable performances in many different problems. Combining Euclidean distance and feature weights, we get the weighted Euclidean distance shown as follows:
Weighted Euclidean Distance

$$\text{DIS}_{AB} = \sqrt{\frac{\sum_{k=1}^{n} w_k \times (A_k - B_k)^2}{\sum_{k=1}^{n} w_k}}, \tag{4}$$

where w_k is the weight of the $k^{\rm th}$ numeric feature. In this study, we use the weighted Euclidean distance function as the similarity measurement for CBR.

4.2 Retrieval Strategy: k-nearest Neighbors

The choice of an effective retrieval strategy is a key element in developing a CBR system. One of the simplest and most straightforward retrieval methods is nearest neighbor (1NN) matching. During generalization, 1NN uses the distance metric to determine the similarity between two cases and predict the output of the query case based on the retrieved nearest case. However, 1NN is highly sensitive to noise. KNN is a more sophisticated approach that can reduce sensitivity to noise and smooth the decision boundaries by setting k greater than

1 (Dasarathy 1991). KNN searches out a neighborhood consisting of k cases that are nearest to the query case and determines the outcome based on the predominance of a particular label in the neighborhood.

Although KNN is quite simple and easy to implement, it performs well in many situations. Cover & Hart (1967) pointed out that the classification error of KNN asymptotically approaches the Bayes error and can also approximate it. Furthermore, Kuramochi & Karypis (2001) found that KNN is even able to outperform an SVM in gene classification using expression profiles, which demonstrates the effectiveness of KNN despite its simplicity.

Determination of the value of k, the number of nearest neighbors, is a key element to consider in CBR. The value of k is related to the similarity measurement and the specific problem, and is typically determined through trial-and-error processes. If k is too high, the retrieval process includes too many insignificant cases, which may lead to poor results, while a small value of k may cause CBR to overlook effective similar cases that would contribute to the correct decision.

4.3 Optimizing Weight Assignment of Numeric Features and Number of Neighbors Using a Genetic Algorithm

Kolodner (1988) suggested using the experience of human experts and statistical evaluations to assign weights to every numeric feature. Experts are expected to have abundant experience and be able to determine reliable features that are more important. However, it is nevertheless difficult for a solar flare forecaster to determine the weights of every image numeric feature. The value of the parameter k, the number of nearest neighbors, is closely related to the similarity measurement and is also difficult for a solar flare forecaster. Hence, a genetic algorithm (GA) is proposed to simultaneously optimize the weights of numeric features and the parameter k, taking into account their mutual relationship.

As a stochastic search technique inspired by ideas from natural genetics and by evolutionary principles, a GA (Drezner & Miseviius 2013) is a powerful and robust method for solving global optimization problems in large and complicated spaces. In particular, the GA can tackle multi-parameter optimization problems with an objective function subject to constraints. In contrast, many traditional searching techniques use a hill-climbing method with an initial value and are prone to falling into sub-optimal situations despite having a higher searching speed. A GA searches a problem space with a population of chromosomes, each of which (described as an individual) encodes optimized parameters. Each individual is assigned a fitness based on its performance. For each generation, the population of the next generation is stochastically selected from the current one based on the fitnesses of individuals. They are also recombined and randomly mutated to evolve toward better solutions. The algorithm terminates when some stopping criterion is met, such as no evolution occurring within several generations. More extensive descriptions of a GA are available in Drezner & Miseviius (2013).

To implement a GA for simultaneous optimization of the weights of numeric features and the parameter k in case of retrieval, we need to specify:

1. The ranges for feature weights and the value of k.
2. An effective and suitable fitness function to evaluate individual performances.
3. Certain parameters of GA, such as data type of the population, population size, crossover rate, mutation rate, etc.

An individual represents the weights of the features extracted from magnetograms and the number k of nearest neighbors. Binary strings are used as the data type for individuals. An example of an individual is illustrated in Figure 3. Since the first n adjustable parameters are coded for weights that reflect the relative importance of each feature, we set the range between 0 and 1. As for the value of k, given the fact that it cannot be too large, the range for k is defined between 1 and 64.

For the fitness function, different evaluation criteria, such as the overall prediction accuracy and the Heidke skill score (HSS), can be used. Different concerns can be associated with different fitness functions. From the perspective of improving the prediction accuracy as a whole, the overall prediction accuracy should be adopted, while from the perspective of improving the predictive power over random forecasting, it is advisable to use the HSS. The overall prediction accuracy and the HSS are explained as follows:

The labels of the samples are grouped into positive (flaring) and negative (non-flaring). The prediction performance is measured by the true positive rate (TP rate) and true negative rate (TN rate).

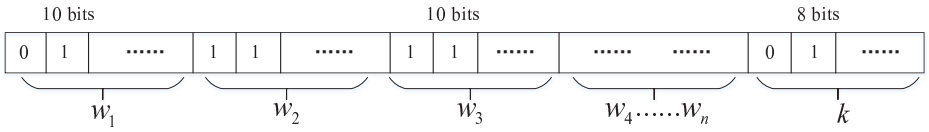

Fig. 3 An example of a coded individual. An individual is binary coded for each weight w of each feature and the number of nearest neighbors k is shown here. The coding represents the fact that an individual contains n features and a value k. The weights w_i of the features are each encoded in 10 bits ranging from decimal values between 0 and 1 and the value of k is encoded in 8 bits ranging from integers 1 to 64.

The TP rate is defined as the ratio of the number of samples correctly predicted as positive to the number of samples that are actually positive

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (5)$$

The TN rate represents the ratio of the number of samples correctly classified as negative to the number of actual negative samples

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (6)$$

Thus, the overall prediction accuracy of the classifier is defined as follows

$$\text{Prediction Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (7)$$

The HSS is generally used to evaluate the performance of the proposed method (Jolliffe & Stephenson 2003)

$$\text{HSS} = \frac{\text{PC} - E}{1 - E}, \quad (8)$$

where

$$N = \text{TP} + \text{TN} + \text{FP} + \text{FN}, \quad \text{PC} = \frac{\text{TP} + \text{TN}}{N}$$

and

$$E = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP})}{N^2} + \frac{(\text{TN} + \text{FP})(\text{TN} + \text{FN})}{N^2}.$$

E is PC for random prediction, so HSS measures the fractional improvement in predictive power of the method over a random forecast.

5 ADAPTATION OF A RETRIEVED SIMILAR PREVIOUS CASE

Adaptation of a retrieved similar previous case is used to revise and adapt the solar flare eruption results for previous similar image cases in order to generate a more desirable solar flare prediction result for a new image case. Programming a computer to implement case adaptation automatically is currently a challenging task. It is recommended that a solar flare forecaster accesses the corresponding original images for the k nearest similar image cases, following the image case retrieval process. Visualization helps the solar flare forecaster to observe and perceive the retrieved image cases more directly and clearly. Based on synthesis of the suggested solar flare prediction result presented by CBR and the experience of the forecaster, more accurate prediction results are expected.

After the real-world operation of CBR to predict a new image case and the prediction result are evaluated by the actual solar flare eruption result, the case and its actual solar flare eruption result are stored in the case library. The whole process for the image-case-based reasoning system for short-term solar flare prediction is illustrated in Figure 4.

6 EXPERIMENTAL RESULTS AND ANALYSIS

6.1 Example Involving a Solar Flare Forecaster

In prediction involving a forecaster, the CBR system firstly acquires initial full-disk magnetograms and extracts the AR from the magnetograms. Then, three image numeric features of the maximum horizontal gradient ($|\nabla_h B_z|_m$), the length of neutral line (L), and the number of singular points (η) are extracted from the image. Next, the three image numeric features are input into the image-case-based reasoning system and 41 similar image cases are retrieved to generate a preliminary solar flare prediction decision by the CBR system, and if necessary, the similar image cases are presented to the forecaster so that his/her experience can be incorporated into the final solar flare prediction decision.

An example shows how interactive processing works, so that a forecaster can modify the result made by CBR whenever the system provides an incorrect solution.

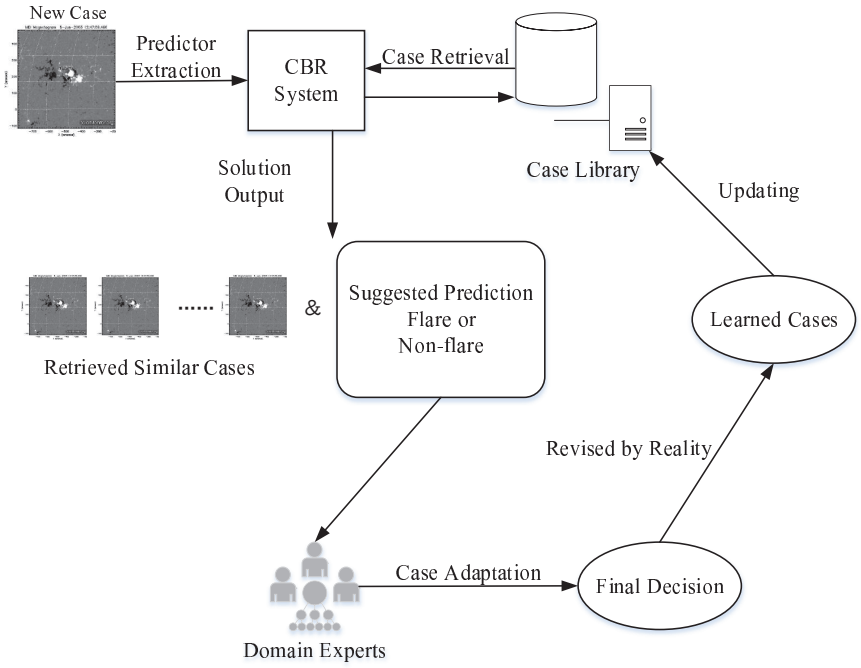

Fig. 4 The whole process illustrating the image-case-based reasoning system for short-term solar flare prediction.

The case "Observation time 23:59:30, 2004 September 8, AR 10669, non-flaring" is incorrectly predicted as flaring. The result of the new case and retrieved similar cases are shown in Table 2.

In this instance, more detailed information will be provided to domain experts (forecasters). Firstly, details of the AR for each retrieved similar case are provided, as shown in Table 3. There are 20 retrieved cases which are non-flaring while 21 cases are flaring, so the corresponding prediction is flaring. However when the analysis report is submitted to a forecaster, it is easy to distinguish between the obfuscating and natural cases. This is because, out of all the retrieved similar cases, there are 19 non-flaring cases from the same event, AR 10409, while 21 flaring cases are dispersed in different ARs. This indicates that the retrieved cases in AR 10409 may share a similar morphological or physical pattern with the new predicted case.

In addition, more detailed information will be submitted to the forecaster. Taking the former case as an example, the information, for the new case and retrieved similar cases, is shown in Table 4. This information consists of observation time, length and width of the AR, folder number of the original image and the label "1" for flaring and "0" for non-flaring. Images (ARs) for the new case and retrieved cases are shown in Figure 5, where the sub-figure (i) refers to the new case and the other sub-figures refer the retrieved cases, arranged from left to right and from top to bottom in correspondence with Table 4. Domain experts can use these detailed pieces of information for each case to help them in making decisions, and so it can facilitate the prediction for the new case. As analyzed above, a forecaster may conclude by modifying the final judgment to non-flaring.

6.2 Solar Flare Prediction Performance of the Image-Case-Based Method in Comparison with Other Techniques

Next, we describe an experiment that was carried out using the whole case library. The samples contained 9801 flaring samples and 45 781 non-flaring ones, which caused a problem with class imbalance. Models derived from unbalanced datasets will be biased toward the class of samples that dominate in terms of quantity. In this experiment, the dataset was undersampled in order to balance the class distribution (Japkowicz & Stephen 2002). The use of random undersampling consists of randomly eliminating elements of the over-sized non-flaring class until it matches the size of the flaring class. The distri-

Table 2 Example of New Case Prediction

New case info.	Number of non-flaring cases	Number of flaring cases
23:59:30, 2004 September 8	20	21
Prediction	Flaring	

Table 3 Statistical Summary of AR for the Retrieved Similar Cases which Incorrectly Predict the New Case

AR	Flaring	Nonflaring	AR	Flaring	Nonflaring
10409	0	19	10508	8	0
10380	4	0	10387	0	1
10528	3	0	10375	6	0

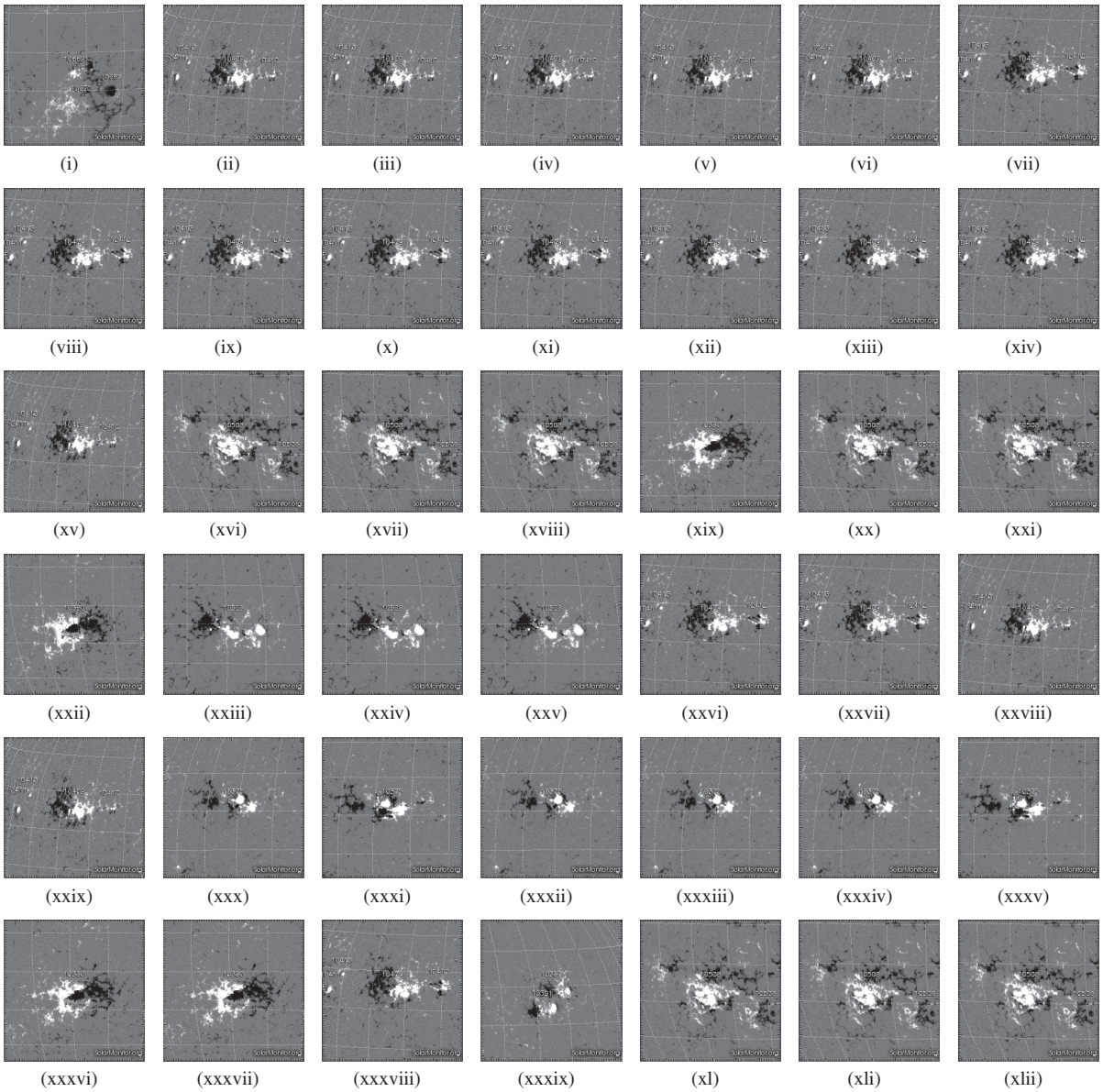

Fig. 5 Images of the new case and retrieved cases. Sub-figure (i) is the new case and other sub-figures are the retrieved cases from *left* to *right* and *top* to *bottom* corresponding to Table 4 accordingly.

Table 4 List of the New Case and Retrieved Similar Cases

AR	Time	AR size: length	AR size: width	Folder No.	Label
10669 (new case)	20040908 23:59:30	104	109	4269	0
10409	20030716 06:27:30	107	180	3848.0004	0
10409	20030716 12:48:30	107	180	3848.0008	0
10409	20030716 08:03:30	107	180	3848.0005	0
10409	20030716 09:36:30	107	180	3848.0006	0
10409	20030716 14:27:30	107	180	3848.0009	0
10409	20030717 03:15:30	107	180	3849.0002	0
10409	20030717 06:24:30	107	180	3849.0004	0
10409	20030717 08:00:30	107	180	3849.0005	0
10409	20030717 09:36:30	107	180	3849.0006	0
10409	20030717 11:12:30	107	180	3849.0007	0
10409	20030717 12:48:30	107	180	3849.0008	0
10409	20030717 17:36:30	107	180	3849.0011	0
10409	20030717 19:12:30	107	180	3849.0012	0
10409	20030716 09:12:30	107	180	3848.0012	0
10508	20031123 03:11:30	111	155	3978.0002	1
10508	20031123 06:27:30	111	155	3978.0004	1
10508	20031123 14:23:30	111	155	3978.0009	1
10380	20030612 03:15:30	111	180	3814.0002	1
10508	20031123 20:51:30	111	155	3978.0013	1
10508	20031123 17:35:30	111	155	3978.0011	1
10380	20030611 23:59:30	111	180	3814	1
10528	20031225 22:27:30	108	201	4010.0014	1
10528	20031225 04:51:30	108	201	4010.0003	1
10528	20031225 08:03:30	108	201	4010.0005	1
10409	20030717 01:39:30	107	180	3849.0001	0
10409	20030717 04:48:30	107	180	3849.0003	0
10409	20030716 16:00:30	107	180	3848.001	0
10409	20030716 22:27:30	107	180	3848.0014	0
10375	20030605 17:36:30	114	166	3807.0011	1
10375	20030606 01:39:30	114	166	3808.0001	1
10375	20030605 06:27:30	114	166	3807.0004	1
10375	20030605 11:15:30	114	166	3807.0007	1
10375	20030605 20:48:30	114	166	3807.0013	1
10375	20030606 08:00:30	114	166	3808.0005	1
10380	20030612 11:11:30	111	180	3814.0007	1
10380	20030612 12:47:30	111	180	3814.0008	1
10409	20030717 00:03:30	107	180	3849	0
10387	20030626 01:39:30	106	64	3828.0001	0
10508	20031123 15:59:30	111	155	3978.001	1
10508	20031123 19:15:30	111	155	3978.0012	1
10508	20031123 23:59:30	111	155	3979	1

bution for the undersampled dataset in the space of three image numeric features is shown in Figure 6. It can be seen that the sizes of the two classes are equal after undersampling. It is also straightforward to conclude that it is difficult to make predictions for samples in the regions where non-flaring and flaring overlap.

Multi-population GA with migrations between populations (Drezner & Miseviius 2013) is employed to boost the possibility of finding the global optimum although it takes more time. HSS is used as the fitness function. Other specific parameters are shown in Table 5. After the optimization process terminates and the best individual is derived, the weights are inserted into the simi-

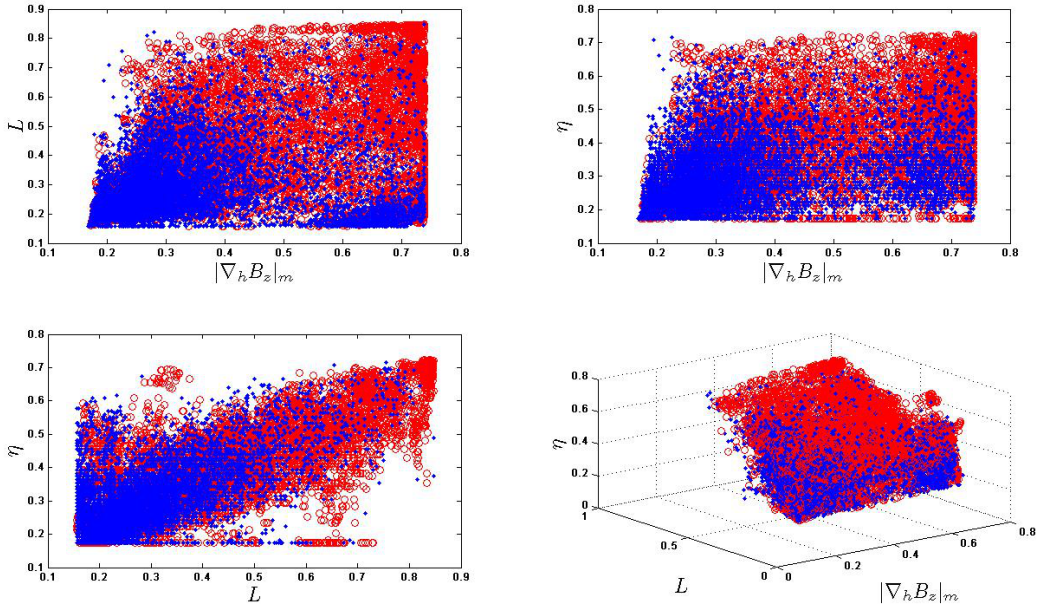

Fig. 6 Distribution of samples in the space of three predictors. The *blue symbols* represent non-flaring samples and the *red symbols* represent flaring samples.

Table 5 Specific Parameters for Multi-population GA

Parameter	Value
Population size	10
Number of populations	10
Preserved elites	2
Migration rate	0.2
Crossover rate	Randomly selected between 0.6 and 0.8 for each population
Mutation rate	Randomly selected between 0 and 0.2 for each population
Termination criterion	Evolution no greater than 1e–6 within 20 generations

larity measurement formula and the number k is inserted into the case retrieval strategy.

In the genetic optimization experiment, the best individual values for the weights of three image numeric features, and for the value of k, are derived simultaneously. The best weights of the three features, $|\nabla_h B_z|_m$, L and η are 0.6211, 0.9221 and 0.1568, respectively, and the derived value of k is 41. With these weights and parameter value, the average HSS and prediction accuracy of CBR are 0.5092 and 75.65%, respectively.

We tested our model with cross validation using ten data groups, in which the samples were from different ARs. The dataset was divided equally into ten folders. Each folder was used in turn as testing data, with the other nine as training data, until every folder had been tested by the others. Hence the final result consisted of a performance average together with a standard deviation. Random undersampling was applied for each training set and test set. As a result, the final sample sizes for two populations within a single folder were equal. The new case and the retrieved cases should not originate from the same AR, primarily because samples from the same AR will, in most cases, share a similar mechanism for chromosphere activity, and hence the earlier samples easily mislead the final judgment of the new case. For example, the first emerging case, AR 10669, needs to be predicted by other cases from different ARs. If we use retrieved cases from the same AR, the prediction for a second case to emerge critically depends on the first case of AR 10669, and therefore, an incorrect prediction for the first case may lead to a prediction failure in the second case. This implies that the basis for inference would be unequal between the two cases. Furthermore, a second (and less significant) factor we consider is that retrieved

cases from the same AR may still be unlabeled, because they are in close proximity to the case that has newly emerged. To avoid these problems, we retrieve similar cases from other ARs.

The Bayesian network, SVM, logistic regression, radial basis function, classification and regression tree, C4.5 decision tree and back-propagation network (Mitchell 1997, Murphy 2012) methods are used in the experiment for comparison with the CBR method. A Bayesian network is a good basis, on which to build probabilistic reasoning models. Bayesian learning algorithms that can be used to calculate explicit probabilities for hypotheses are among the most practical approaches to certain types of learning problems. An SVM is an algorithm that emphasizes the use of structural risk minimization theory. An SVM can operate like a linear model to obtain a description of the nonlinear boundary of a dataset, using a nonlinear mapping transform. Logistic regression is a widely used algorithm for binary classification that implements a logistic transform on the variables being considered. The perceptron learning rule is built using a single hyperplane, with a group of weights assigned to each attribute. Data are classified into one class when the sum of the weights of an attribute is a positive number, and into another class, when the sum is a negative number. In back-propagation theory, attributes are reweighted if the samples are classified incorrectly, until the classification is correct. A radial basis function network is another kind of feed-forward network, which uses a Gaussian function as its activation function and a sigmoid function to transform the classification. The C4.5 decision tree is an extension of the ID3 algorithm. It determines the affiliations of the nodes by using an information gain ratio. C4.5 can discretize continuous attributes, while the ID3 decision tree is restricted to discrete attribute processing. The classification and regression tree is another type of decision tree algorithm that uses the simplicity of the regression method in its binary tree.

In this study, models used for comparison, such as SVM and logistic regression, were hybrid programmed with MATLAB 2010a and the Waikato Environment for Knowledge Analysis (WEKA) platform 3.4.1 which can be freely downloaded from *http://www.cs.waikato.ac.nz/ml/weka/downloading.html*. We utilized a MATLAB script to invoke the WEKA computation kernel, in order to process data in batches. WEKA is a widely used data mining and a machine-learning platform for classification and regression developed by the University of Waikato (Witten et al. 2011) and is advantageous for data preprocessing, parameter optimization and data visualization. Thresholds and parameters were automatically generated by this platform. In particular, the SVM prediction model was coded using the LIBSVM library. The performances of CBR and other techniques using the same dataset are shown as benchmarks for comparison in Table 6. The results demonstrate that CBR offers a slight improvement in accuracy, when compared to other methods. More importantly, the CBR approach can add value by virtue of greater comprehensibility to humans.

6.3 Analysis and Discussion

Among the prediction techniques, the proposed image-case-based reasoning method outperforms some of the benchmark approaches with respect to HSS, TP rate and average accuracy on some data groups. More importantly, it enhances comprehensibility of the prediction results, demonstrating and ensuring the practical applicability of this method in the field of solar flare prediction. In real-world applications, much effort should be devoted to improving the TP rate because if an actual flare was wrongly predicted as a non-flare and no action was taken, then a heavy loss would occur. Therefore, the objective in solar flare prediction is more complex than simply minimizing the misclassification rate. The reasons why the proposed method has the most advantageous performance are analyzed as follows:

First, the image-case-based reasoning method constructs a different local prediction for each individual image case to be predicted, rather than building a global prediction model for the whole case space. There may exist different kinds of patterns of solar flare eruption. Because of this, it is difficult to generate satisfactory prediction results when the prediction model is globally built upon all the cases. Under such circumstances, CBR has an apparent advantage of formulizing less complex local prediction.

Second, the proposed image-case-based reasoning method uses a genetic algorithm to capture and integrate the weight assignments for numeric features and the number of neighbors in case retrieval.

7 CONCLUSIONS

Currently, solar flare prediction mainly focuses on building powerful prediction models and exploring more in-

Table 6 Performance Comparisons between CBR and other Techniques

Method	HSS	TP rate (%)	TN rate (%)	Average accuracy (%)
CBR	0.5092±0.0163	76.38±1.16	74.01±1.59	75.65±0.91
BayesNet	0.4702±0.0189	71.72±1.53	75.31±1.06	73.52±0.94
SVM	0.4706±0.0230	68.17±1.78	78.92±0.87	73.53±1.18
Logistic Regression	0.4582±0.0250	69.14±1.89	76.70±1.16	72.92±1.28
RBF	0.4515±0.0191	69.32±1.97	75.86±1.92	72.58±0.95
Cart	0.4869±0.0188	71.52±1.69	77.18±1.32	74.35±0.95
C4.5	0.4884±0.0210	71.85±2.11	77.01±1.40	74.43±1.06
BP	0.4559±0.0174	62.95±3.02	82.65±2.35	72.80±0.97

formative predictors. However, prediction results provided by some machine learning based models such as neural networks, SVM, etc. are difficult to be understood by forecaster in spite of their possible high prediction accuracy. A common weakness of those models is that they provide a forecaster with little comprehensible information except the final prediction results and the forecaster cannot get involved in prediction to improve the prediction result with their domain of experiences.

Employing image-case-based reasoning for solar flare prediction has two principal advantages over other methods, such as expert systems and model-based prediction using machine learning techniques. The first of these advantages is a better performance according to evaluation criteria such as TP rate, HSS and average accuracy. The second advantage offered by image-case-based reasoning is that it can help forecasters get involved with prediction. Even though it may not be able to offer extreme precision with regard to its overall prediction accuracy, the decision made by the CBR system can provide a useful semi-empirical prediction for a forecaster, who can facilitate their judgment by the use of CBR. Prior information supplied by a forecaster, combined with prediction results from a CBR system, should offer a more satisfactory performance than any other method. Compared to CBR, expert systems are overly dependent on the forecaster or on prior knowledge. On the contrary, model-based methods that use machine learning techniques are able to capture the flare eruption mechanism and its relevant inherent rules adequately, but their reasoning process remains incomprehensible and therefore, the forecaster is unable to offer a complementary suggestion or provide further judgment.

The observational data for solar flare prediction are expressed through images, which provide the most abundant information about flare eruption. Using reasoning with such solar flare image cases is more comprehensible and interpretable, compared to models built on numeric data. Therefore, as described in this article, *SOHO*/MDI longitudinal magnetograms and ARs are used to construct an image-case library, and the image-case-based reasoning method is proposed to predict solar flare eruptions. Genetic optimization algorithms are employed for optimizing the weight assignment of the image features and the number of retrieved similar image cases. Similar image cases and prediction results derived by the majority voting method for similar image cases are output and shown to the forecaster, which can then be integrated with his/her experience to produce the final prediction results. Experimental results demonstrate that the image-case-based reasoning method gets the forecaster more involved in the forecasting process. There is great potential for the image-case-based reasoning method to become a more practical and useful solution for solar flare prediction.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 11078010). We thank the *SOHO*/MDI consortium for the data. *SOHO* is a project of international cooperation between ESA and NASA. We thank the help of Dr. Huang Xin from the National Astronomical Observatories of China for data processing.

References

Abramenko, V. I. 2005, ApJ, 629, 1141

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, Sol. Phys., 283, 157

Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89

Begum, S., Ahmed, M. U., Funk, P., Xiong, N., & Folke, M. 2011, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41, 421

Berry, M. J., & Linoff, G. 1997, Data Mining Techniques: for Marketing, Sales, and Customer Support (John Wiley & Sons, Inc.)

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJ, 747, L41

Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127

Boucheron, L. E., Al-Ghraibah, A., & McAteer, R. T. J. 2015, ApJ, 812, 51

Colak, T., & Qahwaji, R. 2009, Space Weather, 7, S06001

Cover, T., & Hart, P. 1967, IEEE Transactions on Information Theory, 13, 21

Cui, Y., Li, R., Zhang, L., He, Y., & Wang, H. 2006, Sol. Phys., 237, 45

Dasarathy, B. V. 1991, IEEE Transactions on Systems, Man, and Cybernetics, 21, 1140

Drezner, Z., & Miseviius, A. 2013, Computers & Operations Research, 40, 1038

Gallagher, P. T., Moon, Y.-J., & Wang, H. 2002, Sol. Phys., 209, 171

Guo, Y., Peng, Y., & Hu, J. 2013, Computers in Industry, 64, 90

Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175

Huang, X., Zhang, L., Wang, H., & Li, L. 2013, A&A, 549, A127

Japkowicz, N., & Stephen, S. 2002, Intelligent Data Analysis, 6, 429

Jolliffe, I. T., & Stephenson, D. B. 2003, Forecast Verification: a Practitioner's Guide in Atmospheric Science (John Wiley & Sons)

Kolodner, J. L. 1988, Case-based Reasoning: Proceedings of a Workshop on Case-Based Reasoning: Holiday Inn, Clearwater Beach, Florida (Morgan Kaufmann Publishers)

Kuramochi, M., & Karypis, G. 2001, in Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, IEEE, 313

Ledvij, M. 2003, Industrial Physicist, 9, 24

Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277

Leka, K. D., & Barnes, G. 2003b, ApJ, 595, 1296

Li, H., & Sun, J. 2013, Journal of Forecasting, 32, 180

Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Zhan-Le Du 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 441

Liu, Y., Cheng, H., Huang, J., et al. 2012, Journal of Medical Systems, 36, 3975

Mamaghani, F. 2002, Information Systems Management, 19, 13

Marir, F., & Watson, I. 1994, The Knowledge Engineering Review, 9, 355

McIntosh, P. S. 1990, Sol. Phys., 125, 251

Mikos, W. L., Espindola Ferreira, J. C., Gomes, F. G. C., & Lorenzo, R. M. 2010, International Journal of Computer Integrated Manufacturing, 23, 177

Mitchell, T. M. 1997, Machine Learning (McGraw-Hill Education (Asia)), 870

Muranushi, T., Shibayama, T., Muranushi, Y. H., et al. 2015, Space Weather, 13, 778

Murphy, K. P. 2012, Machine Learning: a Probabilistic Perspective (MIT Press)

Osborne, H. R., & Bridge, D. G. 1996, in European Workshop on Advances in Case-Based Reasoning, Springer, 309

Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195

Richter, M. M., & Aamodt, A. 2005, The Knowledge Engineering Review, 20, 203

Riesbeck, C. K., & Schank, R. C. 2013, Inside Case-based Reasoning (Psychology Press)

Shin, S., Lee, J.-Y., Moon, Y.-J., Chu, H., & Park, J. 2016, Sol. Phys., 291, 897

Volobuev, D., Makarenko, N., & Knyazeva, I. 2016, Journal of Physics: Conference Series, 675, 032027 (IOP Publishing)

Wang, J., Shi, Z., Wang, H., & Lue, Y. 1996, ApJ, 456, 861

Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, Advances in Space Research, 42, 1464

Wang, B. X., & Japkowicz, N. 2010, Knowledge and Information Systems, 25, 1

Wheatland, M. S. 2001, Sol. Phys., 203, 87

Wheatland, M. S. 2004, ApJ, 609, 1134

Witten, I. H., Frank, E., & Hall, M. A. 2011, Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Publishers Inc.)

Yu, D., Huang, X., Hu, Q., et al. 2010a, ApJ, 709, 321

Yu, D., Huang, X., Wang, H., et al. 2010b, ApJ, 710, 869

Zhou, X., Stern, R., & Müller, H.c 2012, International Journal of Computer Assisted Radiology and Surgery, 7, 401