The University of Bradford Institutional Repository

Link to original published version: http://dx.doi.org/10.1007/s11207-010-9516-5 (Erratum to this article is available at http://dx.doi.org/x10.1007/s11207-012-0163-x)

Citation: Al-Omari M, Qahwaji RSR, Colak T and Ipson SS (2010) Machine learning-based investigation of the association between CMEs and filaments. Solar Physics. 262(2): 511-539.

Machine Leaning-Based Investigation of the Associations between CMEs and Filaments

M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson

School of Computing, Informatics & Media

University of Bradford, Richmond Road, Bradford BD7 1DP, England, U.K.

(E-mail: m.h.al-omari@brad.ac.uk, r.s.r.qahwaji@brad.ac.uk, t.colak@brad.ac.uk, s.s.ipson@brad.ac.uk)

Abstract. In this work we study the association between eruptive filaments/prominences and coronal mass ejections (CMEs) using machine learning-based algorithms that analyse the solar data available between January 1996 and December 2001. The Support Vector Machine (SVM) learning algorithm is used for the purpose of knowledge extraction from the association results. The aim is to identify patterns of associations that can be represented using SVM learning rules for the subsequent use in near real-time and reliable CME prediction systems. Timing and location data in the NGDC filament catalogue and the SOHO/LASCO CME catalogue are processed to associate filaments with CMEs. In the previous studies which classified CMEs into gradual and impulsive CMEs, the associations were refined based on the CME speed and acceleration. Then the associated pairs were refined manually to increase the accuracy of the training dataset. In the current study, a data-mining system has been created to process and associate filament and CME data, which are arranged in numerical training vectors. Then the data are fed to SVMs to extract the embedded knowledge and provide the learning rules that could have the potential, in the future, to provide automated predictions of CMEs. The features representing the event time (average of the start and end times), duration, type and extent of the filaments are extracted from all the associated and not-associated filaments and converted to a numerical format that is suitable for SVM use. Several validation and verification methods are used on the extracted dataset to determine if CMEs can be predicted solely and efficiently based on the associated filaments. More than 14000 experiments are carried out to optimise the SVM and determine the input features that provide the best performance.

1. Introduction

Coronal mass ejections (CMEs) are one of the most spectacular solar events affecting human activities in space or ground-based communication systems. The Earth environment and geomagnetic activity are affected by an outward flow of ionized solar plasma known as the solar wind (Pick, Lathuillere, and Lilensten, 2001). Geomagnetic storms tend to be correlated with CMEs (Wilson and Hildner, 1984): Therefore, predicting CMEs can be useful in the forecasting of the conditions in the space environment (Webb, 2000).

The first report of a CME in the astronomical literature was made in 1860 but no more appeared until the 1970's (Briand, 2003). The assumption of a cause-and-effect relationship between CMEs and solar flares has driven heated arguments (Cliver and Hudson, 2002). The previous researches on CMEs (Munro *et al.*, 1979); Poland *et al.*, 1981); Yashiro *et al*,. 2005) showed that most CME events have associations with eruptive filaments/prominences and/or solar flares. The exact degree of this association is currently not clear and it is one of the long standing uncertainties in solar research because most of the available studies were carried out on only a few years of data or on limited cases. One of the aims of this work is to provide more insight into this uncertainty. It has been noted from the previous researches that some solar features lack clear definitions, which increases the difficulty of designing automated detection and processing systems. In addition, the recent space missions (*Hinode* and STEREO) are generating massive increases in the amount of solar data available, making the processing of this vast amount of data very challenging.

Webb *et al*. (1998) reported a case study of the association between CMEs, magnetic clouds, and geomagnetic storms and concluded that CMEs are the real link between solar eruptions and space weather activities affecting the Earth. A summary of previous researches on the relationships between CMEs and other solar activities is given in Table 1. As it can be seen from Table 1, CMEs are mostly related to solar flares and eruptive filaments. Although some of the researchers (Moon *et al*., 2002; Qahwaji *et al*., (2008c) studied large datasets or data over long periods of time for correlations, most of the researches has been done on limited and concentrated data in order to draw accurate and meaningful conclusions (Gilbert *et al*., 2000; Subramanian and Dere, 2001). In any case, although different degrees of correlations were concluded by the researchers, they were~~are~~ mainly focused on the relationship among CMEs, filaments, and solar flares~~relationship between CMEs and filaments and solar flares~~. Some researchers concentrated their analysis on the Solar Maximum Mission (SMM) data to draw some conclusions about the solar-cycle dependence of the relation between filament eruptions and CMEs. Webb and Hundhausen (1987) studied 58 CMEs observed in 1980 using the HAO Coronagraph/ Polarimeter on the SMM satellite and compared them with other forms of solar activity (eruptive prominences, ~~H~~ Hα flares, soft X-ray events, and metric type II and IV radio bursts). It was found that 66% of the CMEs were associated with these solar activities. Out of these CMEs, 68% were found to be associated with eruptive prominences, 37% were associated with ~~H~~ Hα flares, 76% were associated with X-ray events, and 32% were associated with radio type II or IV events. Another study of SMM data for 73 CMEs between 1984 and 1986 was reported in St. Cyr and Webb (1991). They found that 76% of the CMEs were associated with eruptive prominences, 26% were associated with ~~H~~ Hα flares and 74% with X-ray events. Srivastava, Gonzalez, and Sawant (1997) studied 14 CMEs observed between March and September 1980 using SMM and concluded that strong association existed between CMEs and coronal holes, eruptive prominences and current sheets. Hori and Culhane (2002) used microwave images from the Nobeyama Radioheliograph to examine 50 prominence eruptions near the solar maximum between 1999 and 2000 and concluded that 92% of the prominence eruptions were associated with CMEs.

Currently, five major CME eruption models exist: the thermal blast model, the dynamo model, the mass loading model, the tether release model, and the tether straining model (Low, 1999b; Klimchuk, 2001; Low, 2001a~~; Low, 2001b~~). The last three are storage-and-release type models, where a slow build-up of magnetic stress occurs before an eruption begins (Aschwanden, 2004). The model which is most directly related to our present work is the mass loading model. The mass loading process during the pre-eruption phase of a CME can be manifested in the form of a growing quiescent or eruptive filament. Mass loading can be associated with prominences, which are extremely dense, contained in a compact volume, and of

Formatted: Font: Italic

chromospheric temperature. Prominences are thought to play a major role in CME eruptions because of their simultaneous appearance, according to the observations reported by Low (1996, 1999a). A crucial criterion for the valid model of CME eruptions is the mass of the prominence and its role in the storage of magnetic energy (Low, Fong, and Fan, 2003, Zhang and Low, 2004).

Machine learning and data mining have not been widely applied to solar data. For the references on these subjects, Qahwaji and Colak (2007) reported a comparison of several learning algorithms for the automated short-term prediction of solar flares. Qahwaji *et al*. (2008c) investigated all the reported flares and CMEs between 01 January 1996 and 31 December 2004 (19164 solar flares and 9297 CMEs) and concluded that 17.4% of the reported solar flares are CME-associated on the basis of timing information. The authors compared the prediction performance using Cascade Correlation Neural Networks (CCNN) and Support Vector Machines (SVM).

Al-Omari *et al*. (2008) and Qahwaji *et al*. (2008a, 2008b) reported large-scale studies looking for associations between CMEs and eruptive filaments/prominences based on their location and timing in the solar cycle . In Al-Omari *et al*. (2008) and Qahwaji *et al*. (2008a) approximately 16% of the filaments in the period from 1 January 1996 and 31 December 2006 were associated with CMEs. The former paper used SVM to extract the knowledge contained in the associated datasets while the latter paper used Radial Basis Function (RBF) networks which are a powerful interpolation technique based on curve fitting that can be efficiently applied to multidimensional space. In RBF networks, learning is achieved when a multi-dimensional surface is found providing optimum separation of multi-dimensional training data.

The Adaptive Boosting algorithm (AdaBoost), described in Freund and Schapire (1997), was used in Qahwaji *et al*. (2008b) for CME prediction. They compared three different boosting algorithms (Real, Gentle, and Modest AdaBoost). Real AdaBoost is the boosting algorithm reported in Schapire and Singer (1999), which is a generalisation of the basic AdaBoost algorithm introduced in Freund and Schapire (1996). Gentle AdaBoost, introduced in Friedman, Hastie, and Tibshirani (2000), is a more robust and stable version of the Real AdaBoost algorithm and performs slightly better than the latter on regular data and considerably better on noisy data (Friedman, Hastie, and Tibshirani, 2000). Modest AdaBoost, described in Vezhnevets and Vezhnevets (2005), can provide better generalization capability and higher resistance to over-fitting compared to the alternative forms of AdaBoost. In addition, Modest AdaBoost, in certain cases, can provide good performance in terms of test error.

Our approach in this work is to use data mining and machine learning techniques, which have not been fully exploited before, to verify the associations between CMEs and filaments and to represent the associations using computer-based learning rules, which can then be used to extract knowledge and to provide off-line predictions.

The current work introduces a computer platform for studying the association between CMEs and f~~F~~ilaments within the context of CME predictions. The aims of this study are to:

1. Investigate if a degree of association exists between erupting filaments and CMEs.
2. Investigate if this association can be represented automatically using computerised learning rules.
3. Provide a future work plan on how the outcomes of this study can be used as a part of more comprehensive work for the automated, near real-time prediction of CMEs.

At the current phase of our research work, the word “prediction” is used as an allegorical expression for the use of computerised learning rules in finding the possibility that a filament will initiate a CME. The expression “prediction performance” is used as a measure of how correct is the rule’s decision that a CME will be initiated or not, compared with the actual CME records.

This paper is organized as follows: Section 2 describes the data catalogues, the association principles and discusses different levels of associations. The creation of the training and testing datasets together with the practical implementation and evaluation of the system using machine learning algorithms are discussed in Section 3. Concluding remarks and recommendations for future work are presented in Section 4.

2. Automated Analysis of Solar Data

2.1. Description of the Data Catalogues

Filament data from publicly available catalogues provided by the National Geophysical Data Centre (NGDC)[1] are used in this study. The NGDC filament catalogue holds records including dates, times, locations, physical properties, types, and active region numbers (NOAA) which have been supplied by many solar observatories around the world that have been tracking eruptive filaments/prominences. A sample of this catalogue is shown in Figure 1.a.

It is important to note that the start and end times of each filament in the catalogue are followed by a qualifier with three levels: D (after), E (before) and U (uncertain). In the catalogue, filaments are classified in 15 types as shown in the first column of Table 2. The second column describes these types and the last column lists the numerical representation for each type, as explained in Section 3.2. Filament types in the catalogue are followed by an “importance” parameter that is based on the type and varies from 0- to 3+. The importance is given according to the greatest extension of the filament before activation, apparent length of surges, or the general activity level of a prominence region. The filament extent mentioned in Figure 1a. is given by the radial extent above the limb in hundredths of solar radius for limb events and it is given by the extent in whole degrees for disk events.

Two main types of filaments/prominences were first introduced by the Menzel-Evans scheme of classification (Menzel and Evans, 1953): (1) filaments originating in the coronal space and (2) filaments originating in the chromosphere. Those originating from above in the coronal space consist of spot prominences (loops and funnels) and non-spot prominences (coronal rain, tree trunks, trees, hedgerows, suspended clouds, and mounds). On the other hand, prominences originating from below in the chromosphere include surges and puffs (spot prominences) and spicules (non-spot prominences). Detailed definitions for the filament types listed in Table 2 can be found in the glossaries provided by the Space Weather Prediction Centre (NOAA)[2] and the Space Environment Information System (SPENVIS)[3].

The data contained in the CME catalogue includes all CMEs manually identified since 1996 in the images from the Large Angle and Spectrometric Coronagraph (LASCO) on board the Solar and Heliospheric Observatory (SOHO)[4], generated and maintained by the Centre for Solar Physics and Space Weather at the Catholic University of

[1] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FILAMENTS/, last access: 2008.

[2] http://www.swpc.noaa.gov/info/glossary.html, last access: 2009.

[3] http://www.spenvis.oma.be/spenvis/help/system/glossary.html, last access: 2009.

[4] http://cdaw.gsfc.nasa.gov/CME_list/, last access: 2008.

America. This catalogue of SOHO data has been constructed in cooperation with the Naval Research Laboratory and the Solar Data Analysis Centre (SDAC) at NASA Goddard Space Flight Centre. This CME catalogue provides details of CME appearances, dates and times, position angles, angular widths, speeds and accelerations as illustrated in Figure 1b.

2.2. Associations

Data from the NGDC and SOHO/LASCO catalogues are analyzed by a C++ computer platform created to automatically associate CMEs with eruptive filaments/prominences. In this work, the 'associations' are defined as the learning rules that can be used in the future as part of an automated system for CME predictions.

The system starts by parsing the CME and filament catalogues. Then a filament is labelled either "*A*", for associated, "*PA*" for possibly-associated filament, or "*NA*" for not-associated. Datasets for the *NA* and *A* filaments are created for the extraction of their properties, which are represented using a numerical format that is suitable for the input to the machine learning algorithms. The *PA* filaments are excluded from the machine training to make the learning performance as accurate as possible.

The associations are determined as discussed in the following four steps:

1. *Time-based associations*. The date and time of every CME are compared with the date and time of every filament (Al-Omari *et al*., 2008; Qahwaji *et al*., 2008a). The association labelling starts with the time-based associations. The CME event time is taken directly from the SOHO/LASCO CME catalogue. However, as most of the filament start and end times are reported in the NGDC filaments catalogue as uncertain, the average of the filament start and end times is taken to be the filament event time (Moon *et al*., 2002). As indicated in Figure 2, the width of the time association window is defined to be 2α minutes. If a CME is not recorded in the interval from α minutes before to α minutes after the filament event time, the filament is labelled *NA*; otherwise, it is labelled *PA* and recorded together with the relevant CMEs. To make the data sampling as homogeneous as possible, the value of α was the same in all our experiments and chosen to be 60 minutes, following Moon *et al*. (2002).
2. *Location-based associations*. The central position angle (CPA) of every CME is compared with the polar position of the centroid of every filament (Al-Omari *et al*., 2008; Qahwaji *et al*., 2008a). In this step, the algorithm analyses the *PA* filaments, identified by step 1, and the corresponding CME candidates. The algorithm defines an association sector on the solar disk within ±30º of the centroid of each *PA* filament as shown in Figure 3. If any of the CME candidates of a *PA* filament has a CPA lying within a filament's association sector, the filament is given the label *A* and recorded together with its associated CME. In the cases where the candidates are halo CMEs, the measurement position angle (MPA) is used instead because there is no CPA for a halo CME. According to Yashiro *et al*. (2004) and Gopalswamy *et al*. (2009), MPA is defined for the CME's leading edge as the position angle at which the height-time information is measured for its fastest moving part. Apart from CMEs that have a non-radial movement, the CPA and MPA are equal (Gopalswamy *et al*., 2009). So, MPA can be used as an indicator of the CPA.

3. *Refining associations based on a CME's speed and acceleration*. According to Sheeley *et al*. (1999), CMEs can be classified into two classes: gradual and impulsive. The gradual CMEs are accelerating, with speeds ranging between 400 to 600 km s^{-1} and are associated with eruptive activities. The impulsive CMEs are decelerating, with speeds faster than 750 km s^{-1} and are initiated by solar flares. It is reported in Moon *et al*. (2002) that the median acceleration and speed for CMEs associated with significant flares (M and X classes) are -8m s^{-2} and 636m s^{-1}, respectively. Such CMEs can be assumed to be impulsive CMEs. By examining the distributions of acceleration and speed of filament-associated CMEs in steps 1 and 2, it is found that these CMEs have zero median acceleration and a median speed of 417.5km s^{-1} as shown in Figure 4. As our algorithm associates CMEs with eruptive filaments/prominences, it is dealing with gradual CMEs (Sheeley *et al*., 1999). We therefore decided to apply strict association conditions that could lead to more accurate knowledge extraction with better machine learning performance. By making a simple comparison between the statistics of gradual and impulsive CMEs in our sample of data and those of Moon *et al*. (2002), it is clear that all CMEs that have accelerations less than -8m/s^2 and speeds greater than 636km s^{-1} are more likely to be associated with significant solar flares. Hence, we refined our associations by ignoring any *A* filament with associated CME having acceleration less than -8m s^{-2} or speed greater than 636km s^{-1}.
4. *Manual refinement*. By examining the association algorithm from the previous three steps, it is apparent that the number of associated filaments might be greater than the number of associated CMEs which means that a single CME could have been associated with more than one filaments. One should also consider the possibility of the data sets including single filaments that are each associated with more than one CMEs. These cases have been dealt with in the following way:
   - If a filament has more than one CME candidates then the algorithm will associate it with the closest CME in time and discard the rest.
   - If the same CME is associated with many filaments then the case is investigated manually using ~~H~~ Hα solar images that are obtained from Meudon Observatory[5] and the Big Bear Solar Observatory (BBSO)[6]. We compare such filaments according to their distance from the limb, angular distance from the CME, duration, and extent. It is assumed that the associated filament is likely to be the one furthest from the centre of the solar disk, nearest to the CME, with longest time duration or alternatively greatest spatial extent.

An example of *PA* filament~~d~~ is given in Figure 5 with its relevant CME. The marked filament started the eruption ~~started~~ at 9:40 and disappeared at 10:15 on 19 July 2001 (the calculated event time is 9:57:30). The CME was first recorded on the same day at 10:30 (about 32 minutes after the filament event time) which falls within the filament time association window. The *PA* filament was centred at S20W59 (a polar angle of 251º) and the CME had a central position angle of 275º which falls within the filament association region. Hence, the filament is labeled as an *A* filament. This example is the case where the disappearing time of disappearing filaments was treated as the end time.

[5] http://bass2000.obspm.fr, last access: 2008.

[6] http://www.bbso.njit.edu/pub/archive/, last access: 2008.

By applying step 1 of the association algorithm, a total of 6101 out of 7332 filaments were classified as *NA* filaments based on their timing information. A total of 1231 filaments were classified as *PA* filaments with 866 CME candidates out of 5449 events recorded in the CME catalogue. The *PA* cases were compared on the basis of their locations and only 465 filaments were re-classified as *A* filaments, together with 330 CME events. Here, it is interesting to note that the association algorithm associated 6.1% of the reported CMEs in the period 1996 to 2001 with filaments. This result is comparable with that obtained by Moon *et al*. (2002) who reported that 4% of the CMEs in the period 1996 to 2000 were associated with filaments on the basis of time and location using the same time-window width of 2 hours. Zhou, Wang, and Cao (2003) reported that more than 94% of halo CMEs in the period from 1997 to 2001 were associated with eruptive prominences/filaments, but it is impossible to compare this result with ours because these authors did not include all available CMEs in the period. Instead they only selected 197 front-side halo CMEs.

After applying the conditions on the distribution of the speed and acceleration of CMEs, which is the third step of the algorithm, we have discarded a total of 121 CME events so that only 209 out of the 5449 CMEs (3.84%) are associated with a new set of 279 *A* filaments. Refining these association results manually, as described previously in step 4, resulted in the final classification from our association algorithm which is 209 *A* cases, 6101 *NA* cases, and 1022 *PA* cases. Here, it is important to mention that the final association dataset contains only 16 halo CMEs (7.7%), where the MPA is used to provide an indicator for CPA.

The location-based association condition (a constant association sector width of 60º) could be unreliable when associating filaments with the CMEs have larger angular widths. For this reason, we checked our algorithm using a dynamic association sector such that the sector width is set to 60º for CMEs with an angular width <60º and it is set to the angular width of the CME under consideration for CMEs with a larger angular width. By applying the association algorithm again we got the same association results as the final classifications mentioned previously plus an extra 21 associated CME events with an angular width >60º. Because of the large angular widths of these extra CMEs, they were associated with many filaments. For example, a partial halo CME was recorded on 19 Oct 1996 at 17:17 with an angular width of 170º and CPA of 159º. This CME was associated with four filament records having the centroid coordinates at S08E47, S09E41, S28E90 and S19E55. After checking ~~H~~ Hα images it was found that these filaments have approximately the same angular distance of about 50º from the CPA of the CME and therefore it is hard to decide which filament is the relevant one. We prefer to exclude the extra 21 cases from the learning part of our study because we believe that having a small dataset of correctly associated CME-filament pairs is better than having a larger dataset that contains some incorrectly associated pairs.

3. Practical Implementation and Results

3.1. Training and Verification Methods

The present study has used SVMs which have proven to be very effective learning algorithms in similar applications (Qahwaji and Colak, 2007; Qahwaji *et al*., 2008c). All the experiments were carried out using the “MySVM” software (Rüping, 2000). The Anova-Kernel SVM has been used as it was found to outperform the NNs used for solar data processing as explained in Qahwaji and Colak (2007). The Anova kernel is defined by the sum of exponential functions in the *x* and *y* directions,

$$k(x,y)=\left[\sum_{i}\exp\left(-\gamma\left(x_i-y_i\right)\right)\right]^{d} \qquad (1)$$

where the parameters *d* (the exponential degree) and *γ* control the shape of the kernel. Optimisation of the SVM performance was done by adjusting *d*, *γ* and the classification threshold. The classification threshold is simply the decision value at which the data can be classified into two classes. Therefore, SVM classification marks above this threshold would be associated with class 1 (which initiates a CME in our work) and the rest of the data would be associated with class 2 (which does not initiate a CME).

The so-called Jack-knife technique is used to provide a correct statistical evaluation of the performance of a classifier when it is trained and tested on a relatively limited number of samples. The technique divides the total number of samples into two sets: a training set and a testing set. In practice, a random number generator is used to divide the samples into training and testing groups. For a finite number of samples, an error counting procedure can be used to estimate the performance of the learning algorithms (Fukunaga, 1990). We did not use the cross validation technique because there are many more negative instances (*NA* filaments) than positive instances (*A* filaments) in our sample of data and the samples were sorted according to the solar cycle timing information which increases the chance that a given subsample may not contain any CME-associated filaments as there are no significant solar activities during the solar minimum; consequently, this will reduce the classifier training performance.

The following performance indicators are used: True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Accuracy, Specificity, Sensitivity, and Heidke skill score (HSS). Since the system is designed to predict if an eruptive filament is going to initiate a CME (positive) or not initiate a CME (negative), we define these indicators as f ollows:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\text{Total actual positives (number of } A \text{ cases)}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (2)$$

where TP (true positives) is the total number of cases for which the system correctly predicts that a filament produces a CME and FN (false negatives) is the number of cases where the system predicts incorrectly that a filament does not produce a CME,

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\text{Total actual negatives (number of NA cases)}}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \qquad (3)$$

where FP (false positives) is the total number of cases for which the system predicts incorrectly that a filament produces a CME and TN (true negatives) is the number of cases where the system predicts correctly that a filament does not produce a CME,

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \qquad (4)$$

where the summation TP+FP+TN+FN is the total number of *A* and *NA* filaments found in our experiments.

Specificity is an indicator of a system’s ability to correctly identify negatives. From Equation (3) and the definition of TN, Specificity = 1−FPR = TNR. Sensitivity, on the other hand, is an indicator of a system’s ability to correctly identify positives and can be defined as the ratio of the number of true positives to the sum of true positives and false negatives or in other words, Sensitivity = TPR.

The Heidke skill score is reported in Heidke (1926) and Balch (2008) and defined as

$$\mathrm{HSS} = \frac{\mathrm{TP} + \mathrm{TN} - E}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FP} - E} \qquad (5)$$

where E is the number of correct predictions which would be made by chance and is calculated as

$$E = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FP} + \mathrm{TN})(\mathrm{FN} + \mathrm{TN})}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \qquad (6)$$

Here HSS ranges from -1 (which means all incorrect prediction) to +1 (which means all correct prediction). If a prediction system has zero HSS, then the system performance is no better than that from random guessing (Balch, 2008).

All of these indicators were calculated when testing the prediction system while using the learning rules which have been extracted from associations. The performance of the system was evaluated using Receiver Operating Characteristic (ROC) curves, as explained in Fawcett (2006). An ROC curve plots FPR on the *x*-axis and the corresponding TPR on the *y*-axis such that the diagonal line corresponds to random guessing (Fawcett, 2006). According to Fawcett (2006), the system with best performance is the one in the ROC curves which is furthest from the diagonal line in the upper-left direction. Mathematically, if we have different systems/configurations and each one is represented on an ROC curve by a point (FPR,TPR) then the system/configuration point with the maximum distance to the diagonal line, in the upper-left direction, has the best performance. The distance D_{ROC} from a point (FPR_i,TPR_i) to the diagonal line can be expressed as:

$$D_{\mathrm{ROC}} = \frac{|\mathrm{FPR} - \mathrm{TPR}|}{\sqrt{2}}. \qquad (7)$$

3.2. Data Handling

Numerical representations were used for *A* and *NA* filaments as machine learning algorithms deal mainly with numbers. Properties such as starting time, ending time, type and spatial extent of the filaments can be extracted from the NGDC filament catalogue. Initially we also considered including other properties such as filament location, orientation, and importance. However, unfortunately the necessary data are not provided for a large proportion of the associated filaments and the only location indicator that is available for all filaments is the centroid location. For example, about 63% (4606 out of 7332) of the filament records, for the period 1996-2001, are reported without importance. The lack of data made it impossible for us to use the importance for our experiments. Hence, we decided to use only the groups of properties shown in Table 3.

The timing in Table 3 represents the Julian date of the filaments. As explained before, the event time for a filament was considered to be the average of the start and end times. The Julian date was calculated and normalised to be in the range between 0.1 and 0.9 for the time period to be spanned by the timing input. The distriburion of filaments according to the solar cycle phase is shown in Figure 6 for both *A* and *NA* filaments. The filament duration was calculated as the time difference in hours between the end and start times and then it was normalised between 0.1 and 0.9. The distributions of duration for *A* and *NA* filaments are shown in Figure 7. As mentioned previously, the filament extent is measured in different ways for disk and limb events. Therefore, each input group having information on filament extent was divided into two groups: one for disk events (4D, 3aD, 3bD and 2bD) and one for limb events (4L, 3aL, 3bL and 2bL). Then, the filament extent was normalized in the range from 0.1 to 0.9 for disk and limb events separately and its distribution is depicted in Figure 8. For the filament type parameter to have a meaningful numerical value it can be represented by its probability within the associated filaments and this probability can be calculated from the distribution of filament types of Figure 9. However, some types have almost equal numbers of associated filaments such as the DSD and APR events and other types are not associated with CMEs like the types CAP, CRN, MDP, and SSB. In such cases, it will be impossible for the SVM classifier to distinguish between different types of filaments because they are represented by values that are not separated enough for successful learning and output class separation. Hence, we have decided to represent the filament classes in numerical codes for our learning experiments as listed in Table 2. It is important to point out that these numerical values are no more than codes assigned to each class; they are neither weights nor represent the probability distribution of these classes. Finally, the target function for the input groups is represented by two values: 0.9 indicates a filament initiating a CME and 0.1 indicates a filament not initiating a CME.

3.3. Validation Methods

In the previous works (Al-Omari *et al*., 2008; Qahwaji *et al*., 2008a) it was found that group 3 and group 2a were the best input groups in the context of CME association and prediction. Nevertheless, we decided to carry out another extensive set of experiments attempting to increase the accuracy of our prediction system and to determine the significance of each property within this context.

We created training datasets including 40% *A* filaments and 60% *NA* filaments. Training and testing experiments were carried out and the prediction performance was evaluated using the following two validation methods.

3.3.1. Validation Method 1

The machine learning/training and testing experiments in the first method were carried out with the aid of the Jack-knife technique (Fukunaga, 1990). This is done using 80% of randomly selected samples for training to find the best parameters and topologies for the learning algorithms and the remaining 20% for testing. As mentioned previously, our learning dataset contains 209 *A* filaments which represent 40% of the dataset and we randomly selected another 313 *NA* filaments (60%) to build a complete dataset of 522 filaments. A total of 418 associated and not-associated filaments were used for training. This constituted 80% of the total number of cases. The remaining 104 associated and not-associated filaments were used for testing. The above numbers apply for all the input groups that have no information on filament extent (groups 3, 2 and 2a). Unfortunately, filament extent is not always reported in the NGDC catalogue so that we discarded the associated filaments having no information on their extent, from the training and testing datasets in groups 4D, 4L, 3aD, 3aL, 3bD, 3bL, 2bD and 2bL. In these cases, the number of *A* filaments is reduced to 143 (117 for disk events and 26 for limb events), which means there are only 175 and 39 *NA* filaments for disk and limb events, respectively. Hence, for disk events (groups 4D, 3aD, 3bD and 2bD) we have a dataset of 292 associated and not-associated filaments (234 for training and 58 for testing). And for limb events (groups 4L, 3aL, 3bL and 2bL) we have a dataset of 65 associated and not-associated filaments (52 for training and 13 for testing).

3.3.2. Validation Method 2

In the second validation method, we tried to measure the ability of our system to constitute a near real-time automated CME prediction system. Therefore, we decided to validate our system on some arbitrary selected years of data without the need for random sampling of data using the Jack-knife technique. In this work we carried out extensive experiments using six years of data from 1996 to 2001. Here we used the data from years 1996, 1997, 2000 and 2001 for training and the years 1998 and 1999 for testing. We created a training dataset consisting of 149 *A* filaments and 223 *NA* filaments. The testing stage was more challenging because the testing dataset included all 1765 filaments reported in the NGDC catalogue for years 1998 and 1999. Again, because some filaments are reported without information on their spatial extent, the training and testing datasets were reduced while working with input groups 4, 3a, 3b and 2b. For training, a total of 265 filaments were used, consisting of 106 *A* and 159 *NA* filaments. The number of filaments used for testing was reduced to 1504.

3.4. Optimisation and Results

For both validation methods, the performance of the Anova-Kernel SVM was optimised by adjusting the values of degree (*d*), *γ*, and classification threshold. In the optimisation process, the values of *γ* and *d* were both varied from 1 to 10 in steps of 1. In all experiments, the classifier threshold was initialized to the mean of the predicted scores. The optimisation process was applied to the input features corresponding to each of the seven groups shown in Table 3.

3.4.1. Results for Method 1

In the Jack-knife validation method, for each of 100 configurations and 11 input groups, ten experiments were carried out using the Jack-knife technique and the average TPR and FPR values recorded. Hence, 11000 experiments were carried out with 1100 SVM configurations, so 1100 average values of TPR and FPR were produced. To find the optimum SVM system (optimum *d*, *γ* and input configuration), the results were analysed using the ROC analysis technique and are plotted in Figure 10. The system with best performance is the one in the ROC curves which is furthest from the diagonal line in the upper-left direction. The diagonal line corresponds to random guessing (Fawcett, 2006). The best performing SVM configurations can be seen in Figure 11, which is the magnified region labelled Z in Figure 10.

In order to find the classification thresholds that provide the best prediction for the optimum SVM topologies, the threshold values were changed from 0 to 1 in steps of 0.01 for every input feature set and their selected optimum topologies. Then for each threshold value, ten experiments were carried out using the Jack-knife technique and the averages for all performance indicators, defined previously, were calculated. The results of these experiments are summarized in Table 4 and depicted in the ROC curve of Figure 12.

The optimum threshold values were found by choosing the threshold value with the system performance closest to the upper-left corner in the ROC curve. This is seen clearly in Figure 13, which shows a magnified view of the region labelled Z in Figure 12.

As can be seen by inspection of Figures 11 and 13, an SVM classifier that accepts three inputs (group 3) with *d* and *γ* values of 2 and 8 respectively and a classification threshold value of 0.57 provides the best prediction performance. From Table 4, this SVM configuration provides:

- Average TPR and FPR values of 0.65 and 0.22, respectively, which are seen from inspection of Figure 14 to provide better CME prediction performance than that obtained in our previous work: Al-Omari *et al.* (2008) using SVM, Qahwaji *et al.* (2008a) using RBFs, Qahwaji *et al.* (2008b) using the Real and Modest AdaBoost, and Qahwaji *et al.* (2008c) using CCNN. It is clear from the ROC curve of Figure 14 that the best prediction performance using SVM in Qahwaji *et al.* (2008c) has a better TPR value than the current work as it provided TPR value of 0.73, but with a high FPR value of 0.53. On the other hand, a more conservative performance was provided by the Gentle AdaBoost presented in Qahwaji *et al.* (2008b), with TPR and FPR values of 0.46 and 0.12 respectively. The Gentle AdaBoost (Qahwaji *et al.*, 2008b) is better used as a rejection classifier as it makes fewer false alarms.
- An average accuracy of 73% which is the highest accuracy achieved so far in our research on predicting CMEs.
- An average HSS of 0.43, which is significantly better than random guessing. This value indicates that our system has forecasting ability and we are confident that our system is not predicting by chance or because of the statistical distribution of the selected data sample.
- A specificity (or TNR) of 78% which means a useful prediction performance if used as a rejection classifier to predict when CMEs are not likely to occur. We have achieved a specificity of 88% using the Gentle AdaBoost in Qahwaji *et al.* (2008b) but with a low TPR of 0.46. Therefore, with an accuracy of 73% and specificity of 78% it is seen that our current system will be efficient if used as either a positive or a negative classifier tool for the purpose of CME prediction.

The next best performance is achieved by using two inputs (group 2a) with d, γ and threshold values of 10, 7 and 0.55, respectively. This SVM configuration provides TPR, FPR, specificity, accuracy, and HSS of 0.62, 0.24, 76%, 70% and 0.38, respectively. The use of input group 4D provided good results but with lower accuracy and HSS values of 64% and 0.33 respectively.

These results support the findings in Al-Omari *et al.* (2008), Qahwaji *et al.* (2008a, 2008b, 2008c) and it is clear that an increase in the prediction rate has been achieved with the use of more discriminative input features, such as filament type, for the input groups of Table 3.

To draw an accurate conclusion on the importance of filament properties in CME prediction, the same dataset size must be used during validation. Therefore, further experiments were carried out using the same datasets used before for input groups 4D, 4L, 3aD, 3aL, 3bD and 3bL except that the extent property was discarded from these datasets. For comparison purposes, the groups were relabelled as 4D′, 4L′, 3aD′, 3aL′, 3bD′ and 3bL′. Validation method 1 was used and the optimum results of the experiments are summarized in Table 5.

By comparing the values TPR, FPR, accuracy and HSS of groups 4D and 4L in Table 4 with those of groups 4D′ and 4L′ in Table 5 it is clear that discarding the filament extent from the inputs enhanced the prediction performance by reducing its FPR and increasing its accuracy and HSS. By doing the same comparison between the optimum results of groups 3aD, 3aL, 3bD and 3bL in Table 4 and groups 3aD′, 3aL′, 3bD′ and 3bL′ in Table 5 we can conclude that the filament type and duration, particularly the former, are more important indicators for CME prediction than the filament extent. This conclusion supports the findings of some researchers who reported high associations between CMEs and filaments as they considered selected filament types only. An example on this is the study reported in Pojoga and Huang (2003), where the authors considered three classes of sudden disappearances: eruptive, quasi-eruptive and vanishing (thermal disappearances) filaments. They found that 70% of the eruptive filaments were associated with CMEs, while the correlations were weaker for quasi-eruptive and vanishing filaments. Hence, the filament type could be a strong indicator for the possibility of initiating a CME.

A physical explanation for our findings of a strong relationship between the filament types and CMEs can be concluded from the Menzel-Evans classification (Menzel and Evans, 1953) where a filament/prominence is classified based on its material motion (upward or downward), its association with sunspots, and its shape. From Figure 9 it is found that filaments with DSF, EPL and BSL types accounted for about 53.8% of the CME-associated filaments and these types of filaments ascend from the Sun in their initial phase (Menzel and Jones, 1962). In addition, types like ASR (which rise above the limb) and BSD (which emanate from the chromosphere) accounted for 12.9% of the CME-associated filaments. Hence, we conclude that filaments/ prominences that originate from the chromosphere (moving outward) are most likely to be associated with CMEs. On the other hand, it is reported that a loop prominence system (LPS) may appear as a flare in its initial phases (Jones, 1958) and the material in LPS prominences typically originates near the top of the loop and flows downward to the Sun. Our association algorithm managed to associate only 2 LPS prominences with CMEs which suggests that filaments originating in the coronal space (moving downward) are not likely to be associated with CMEs.

Munro *et al*. (1979) studied the CME associations with several forms of solar activity in the period from May 1973 to February 1974. They found that 50% of the CMEs were associated with EPLs solely (without solar flares) and more than 70% were associated with events including EPLs, LPSs, DSFs, DSDs, BSDs and BSLs (with and without flares). In Gilbert *et al*. (2000), an eruptive prominence (EP) is defined as the prominence in which all or part of its material escapes the solar gravitational field. On the other hand, an active prominence (AP) is defined as the prominence showing motion in Hα images with no part of its material escaping the solar gravitational field. Other types of prominences such as sprays (SPY), surges (BSD, DSD, ASR, BSL), explosions, and coronal rain (CRN) were defined in Zirin (1966). Gilbert *et al*. (2000) studied 26 APs (including Zirin's (1966) surges), 18 EPs and 10 DSFs and they found that 94% of the EPs, 46% of the APs and 70% of the DSFs were associated with CMEs. In their classification scheme, Zirin's (1966) sprays and explosions were considered as either EPs or APs. Webb and Hundhausen (1987) studied the CME associations with all Hα eruptive events over the period from March to August 1980 and found that 68% of the CMEs were associated with EPL, DSF, BSL and SPY events. These results support our findings depicted in Figure 9.

All types of filaments/prominences occurring during solar cycle 18 (started in 1944 and ended in 1954) were investigated by Menzel and Jones (1962) who found that filaments/prominences originating in the coronal space (moving downward) represented 93.1% of the recorded prominences. This explains the low associations between CMEs and filaments in our findings and supports our conclusion that the direction of the material motion (upward or downward) of filaments can be used as an indicator for its association with CMEs.

3.4.2.Results for Method 2

In the second validation method, a total of 100 experiments were carried out for each input group and the values of TPR and FPR were used to create the ROC curve shown in Figure 15 from which the optimum SVM configurations were found. To achieve the best performance of our prediction system we varied the value of the classifier threshold from 0 to 1 in steps of 0.01. The values of TPR and FPR for all thresholds and for all inputs groups were used to create the graph of Figure 16 and all the performance indicators were calculated and summarized in Table 6.

From Figure 16 and Table 6 it is clear that the best performance was obtained while using group 3 with *d*, *γ*, and classification threshold values of 6, 2, and 0.64, respectively. This SVM configuration provides TPR, FPR, Specificity, Accuracy, and HSS values of 0.64, 0.18, 82%, 81% and 0.18, respectively. It is shown in Figure 14 that the current work with validation method 2 has better performance compared to the first method using the Jack-knife technique. We believe that our system is the first to use SVM to predict if a CME is likely to be initiated with an accuracy of 81% and at the same time to predict when CMEs are not likely to occur with a specificity of 82%. Again, the next best performance was obtained with group 2a with *d*, *γ*, and classification threshold values of 2, 1, and 0.72, respectively. This configuration provides TPR, FPR, specificity, accuracy and HSS of 0.62, 0.21, 79%, 78% and 0.15, respectively. Better TPR and HSS values of 0.71 and 0.30 were obtained with group 3bL but with lower accuracy of 71%.

From the results of both validation methods, it is clear that the CME prediction performance has been improved compared to our previous work. Checking some of the association cases manually (using ~~H~~ Hα images) and considering the mass loading model for the CME initiation (conditions related to the distributions of the speed and acceleration of CMEs) enabled the association sets to be refined and hence eliminated some of the instances that might be false associations, which produced some improvement in the prediction performance.

4. Conclusions and Future Research

In this work, we have proposed a novel machine-learning-based system that has been trained and tested using six years of data in the NGDC filament catalogue and the SOHO LASCO CME catalogue. The system associates CMEs with filaments and represents these associations numerically in training vectors that are fed to SVM learning algorithms. An optimisation process was applied to the SVM before the learning process was started. The SVM learning algorithm was chosen because of its outstanding classification performance as reported in Qahwaji and Colak (2007) and Qahwaji *et al*. (2008c).

To determine the optimum configuration for the SVM classification system used in this work, many experiments were carried out changing the parameter values *γ* and degree (*d*). Different classification thresholds were tested to determine the optimum configuration using the ROC curves. These experiments used several validation techniques, such as the Jack-knife technique, as described in Section 3.2.

All the reported filaments and CMEs between 1 January 1996 and 31 December 2001 have been investigated. From 5449 CMEs reported in this period, the association software has searched for CME candidates for 7332 eruptive filaments/prominences. For a CME to be associated with a filament it must pass all the following strict conditions: (1) the CME candidate must be initiated within a two-hour interval centred on the filament event time, (2) the time-associated CME must be located within ±30º of the filament's centroid, (3) this CME must have an acceleration greater than -8m/s^2 and (4) it has a speed less than 636km s^{-1}. Applying these conditions, the algorithm found 209 CMEs (3.84% of the total) to be associated with 279 filaments. The association results were refined manually to remove any repeated associations.

After determining the optimum configurations for the SVM using the Jack-knife technique, the best CME prediction performance for the feature sets considered achieved average TPR, FPR, and TNR values of 0.65, 0.22, and 0.78, respectively. This is a good result as it corresponds to an average accuracy of 73% and a Heidke skill score of 0.43. Further training and validations were carried out by training the system on data from 1996, 1997, 2000 and 2001 and testing the performance on data from 1998 and 1999. For this data, the system achieved average TPR, FPR, TNR, and accuracy values of 0.64, 0.18, 0.82, and 81%, respectively.

In other words, if we use the information from the observed filament (solar cycle time, duration, and type) as an input to our system, the system can predict if this filament is going to initiate a CME with a true positive prediction probability of 65%. At the same time, the system can predict if there will be no CME initiated by the input filament with a true negative prediction probability of up to 82%. Therefore, the whole system, when used for predicting CMEs, can achieve a correct prediction probability of 73%.

It is found that an increase in the accuracy of association/prediction has been achieved with the use of more discriminative features such as the filament type. In the final association results, about 66.7% of the CME-associated filaments are found to be emanating from the chromosphere or moving outward. We conclude that filaments/prominences that originate from the chromosphere (moving outward) are most likely to be associated with CMEs, while filaments originating in the coronal space (moving downward) are not likely to be associated with CMEs.

We believe that this work is important because for the first time the association between filaments and CMEs has been explored and verified using machine learning. This association has been represented using computerised learning rules. As discussed in Qahwaji *et al.* (2008c) this representation is an important step for creating automated and reliable prediction systems that can predict the extremes of space weather. For our system to be near real-time, the detection and classification system for the filaments, mentioned in Figure 17, is needed and it is going to be part of our future work. However our work is far from complete and the prediction performance is not as high as it should be because of the following circumstances that still need to be addressed:

- A large number of filaments are missing from the NGDC filament catalogue. This has been deduced by comparing the data in the filament catalogue with the synoptic maps produced by the Meudon Observatory, which are available publicly at http://bass2000.obspm.fr. The number of filaments reported in the catalogue for years 1996, 1997, 1998, 1999, 2000 and 2001 are 1989, 2506, 1320, 446, 593 and 479, respectively. It is clear that there are many data discrepancies including missing and repeated features. This problem clearly affected our findings as the lost data in years 2000 and 2001 will bias our learning-rule-based SVM system to predict incorrectly that filaments within this period are more likely not to initiate CMEs.
- CMEs can be associated with erupting filaments/prominences and solar flares. However, in this study, only CME associations with filaments were considered and solar flare associations produced in the previous work (Qahwaji *et al.*, 2008c) are not considered. To enhance the CME prediction accuracy it is necessary to combine both association algorithms. This will be investigated in the near future.
- The current work does not distinguish between the front side and backside CMEs and it is possible for the present system to associate a filament with a

backside CME. For example, our association algorithm has managed to associate a CME-filament pair on 30 June 1999 where the CME event was recorded at 13:31 and the filament was first observed at 12:55. However, it is reported in the preliminary list[7] of CME events, which is generated by the LASCO team, that this CME event is a partial halo backside event. The association algorithms have used most of the data reported in the catalogues without the use of solar images. There is only a small difference in the visibility of front side and backside CMEs, so it is very hard to distinguish them using only coronagraph observations (Yashiro *et al*., 2006). It would be desirable to confirm that a CME originates from the front side by checking the images of the lower corona obtained by the Soft X-ray Telescope (SXT) on *Yohkoh* and the Extreme ultraviolet Imaging Telescope (EIT) on SOHO. This will be investigated in future work.

Formatted: Font: Italic

References

Al-Omari, M., Qahwaji, R., Colak, T., Ipson, S.: 2008, In: Saleem, A.I., Barakat, S. (eds.), *5th International Multi-Conference on Systems, Signals and Devices (IEEE SSD 2008)*. 1.

Aschwanden, M.J.: 2004, Physics of the Solar Corona: An Introduction, Praxis Publishing, Chichester, UK, p. 704~~<page>.~~

Balch, C.C.: 2008, *Space Weather* **6**, S01001.

Briand, C.: 2003, *Astron. Nachr*. **324**, 357.

Cliver, E.W., Hudson, H.S.: 2002, *J. Atmos. Solar-Terr. Phys*. **64**, 231.

Fawcett, T.: 2006, *Pattern Recog. Lett*. **27**, 861.

Freund, Y., Schapire, R.E.: 1996, In: Blum, A., Kearns, M. (eds.), *Proc. Ninth Annual Conference on Computational Learning Theory*, 325.

Freund, Y., Schapire, R.E.: 1997, *J. Comp. System Sci*. **55**, 119.

Friedman, J., Hastie, T., Tibshirani, R.: 2000, *Ann. Stat.* **38**, 337.

Fukunaga, K.: 1990, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, p. 220.~~<page>.~~

Gilbert, H.R., Holzer, T.E., Burkepile, J.T., Hundhausen, A.J.: 2000, *Astrophys. J.* **537**, 503.

Gopalswamy, N., Shimojo, M., Lu, W., Yashiro, S., Shibasaki, K., Howard, R.A.: 2003, *Astrophys. J.* **586**, 562.

Gopalswamy, N., Yashiro, S., Michalek, G., Stenborg, G., Vourlidas, A., Freeland, S., Howard, R.: 2009, *Earth, Moon, Planets* **104**, 295.

Heidke, P.: 1926, *Geograf. Ann.* **8**, 301.

Hori, K., Culhane, J.L.: 2002, *Astron. Astrophys.* **382**, 666.

Jing, J.: 2005, 'Dynamics of Filaments, Flares and Coronal Mass Ejections (CMEs)', Ph. D. Thesis, State University of New Jersey, Newark, New Jersey.

Jing, J., Yang, G., Wang, H.M.: 2003, *Bull. Am. Astron. Soc*. **35**, 815.

Jing, J., Yurchyshyn, V.B., Yang, G., Xu, Y., Wang, H.: 2004, *Astrophys. J.* **614**, 1054.

Jones, F.S.: 1958, *J. Roy. Astron. Soc. Canada* **52**, 149.

Klimchuk, J.A.: 2001, In: Song, P., Singer, H., Siscoe, G. (eds.), *Space Weather, AGU Geophys. Monogr*. **125**, 143.

Low, B.C.: 1996, *Solar Phys*. **167**, 217.

Low, B.C.: 1999a, In: Habbal, S.R., Esser, R., Hollweg, J.V., Isenberg, P.A. (eds.), *Solar Wind Nine, AIP Conf. Proc.* **471**, 109.

Low, B.C.: 1999b, In: Brown, M.R., Canfield, R.C., Pevtsov, A.A. (eds.), *Magnetic Helicity in Space and Laboratory Plasmas, AGU Geophys. Monogr*. **111**, 25.

Low, B.C.: 2001a, *J. Geophys. Res.* **106**, 25141.

~~Low, B.C.: 2001b, In: Murdin, P. (ed.), *Encyclopedia of Astronomy and Astrophysics*, Institute of Physics Publishing, Bristol, <page>.~~

Low, B.C., Fong, B., Fan, Y.: 2003, *Astrophys. J.* **594**, 1060.

Menzel, D.H., Evans, J.W.: 1953, *Accademia Nazionale dei Lincei* **11**, 119.

[7] http://lasco-www.nrl.navy.mil/index.php?p=content/cmelist, last access: 2009.

Menzel, D.H., Jones, F.S.: 1962, *J. Roy. Astron. Soc. Canada* **53**, 193.

Moon, Y.J., Choe, G.S., Wang, H., Park, Y.D., Gopalswamy, N., Yang, G., Yashiro, S.: 2002, *Astrophys. J.* **581**, 694.

Munro, R.H., Gosling, J.T., Hildner, E., MacQueen, R.M., Poland, A.I., Ross, C.L.: 1979, *Solar Phys.* **61**, 201.

Pick, M., Lathuillere, C., Lilensten, J.: 2001, *ESA Space Weather Programme Feasibility Studies*, Alcatel-LPCE Consortium.

Pojoga, S., Huang, T.S.: 2003, *Adv. Space Res.* **32**, 2641.

Poland, A.I., Howard, R.A., Koomen, M.J., Michels, D.J., Sheeley, N.R.: 1981, *Solar Phys.* **69**, 169.

Qahwaji, R., Al-Omari, M., Colak, T., Ipson, S.: 2008a, In: Villanueva, J.J. (ed.), *IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2008)*, ACTA Press, 808.

Qahwaji, R., Al-Omari, M., Colak, T., Ipson, S.: 2008b, In: Mahasneh, J. (ed.), *Mosharaka International Conference on Communications, Computers and Applications (MIC-CCA 2008)*, Mosharaka for Researches and Studies, 37.

Qahwaji, R., Colak, T.: 2007, *Solar Phys.* **241**, 195.

Qahwaji, R., Colak, T., Al-Omari, M., Ipson, S.: 2008c, *Solar Phys.* **248**, 471.

Rüping, S.: 2000, *MySVM Manual, Lehrstuhl Informatik* **8**, University of Dortmund,.

Schapire, R.E., Singer, Y.: 1999, *Machine Learning* **37**, 297.

Sheeley, N.R., Walters, J.H., Wang, Y.-M., Howard, R.A.: 1999, *J. Geophys. Res.* **104**, 24739.

Srivastava, N., Gonzalez, W.D., Sawant, H.S.: 1997, *Adv. Space Res.* **20**, 2355.

St. Cyr, O.C., Burkepile, J.T., Hundhausen, A.J., Lecinski, A.R.: 1999, *J. Geophys. Res.* **104**, 12493.

St. Cyr, O.C., Webb, D.F.: 1991, *Solar Phys.* **136**, 379.

Subramanian, P., Dere, K.P.: 2001, *Astrophys. J.* **561**, 372.

Vezhnevets, A., Vezhnevets, V.: 2005, *Modest Adaboost – Teaching Adaboost to Generalize Better*, Graphicon.

Webb, D.F.: 2000, *J. Atmos. Solar-Terr. Phys.* **62**, 1415.

Webb, D.F., Cliver, E.W., Gopalswamy, N., Hudson, H.S., St. Cyr, O.C.: 1998, *Geophys. Res. Lett.* **25**, 2469.

Webb, D.F., Hundhausen, A.J.: 1987, *Solar Phys.* **108**, 383.

Wilson, R.M., Hildner, E.: 1984, *Solar Phys.* **91**, 169.

Yang, G., Wang, H.: 2002, In: Wang, H., Xu, R. (eds.), *Solar-Terrestrial Magnetic Activity and Space Environment, COSPAR Colloq.* **14**, 113.

Yashiro, S., Gopalswamy, N., Akiyama, S., Howard, R.A.: 2006, 36th COSPAR Scientific Assembly, 1778.

Yashiro, S., Gopalswamy, N., Akiyama, S., Michalek, G., Howard, R.A.: 2005, *J. Geophys. Res.* **110**, A12S05.

Yashiro, S., Gopalswamy, N., Michalek, G., St.Cyr, O.C., Plunkett, S.P., Rich, N.B., Howard, R.A.: 2004, *J. Geophys. Res.* **109**, A07105.

Zhang, M., Low, B.C.: 2004, *Astrophys. J.* **600**, 1043.

Zhou, G., Wang, J., Cao, Z.: 2003, *Astron. Astrophys.* **397**, 1057.

Zirin, H.: 1966, *The Solar Atmosphere*, Waltham, Blaisdell-Ginn, P. 297~~<page>.~~

Figure 1 (a) NGDC filament catalogue, (b) SOHO/LASCO CME catalogue.

Figure 2 Time-based association between a CME and a filament.

Figure 3 Location-based association between a CME and a filament.

Figure 4 Distributions of speed and acceleration for filament-associated CMEs.

Figure 5 An example of CME-filament association.

Figure 6 Solar-cycle timing distributions for CME-associated and not-associated filaments.

Figure 7 Duration distributions for CME-associated and not-associated filaments.

Figure 8 Extent distributions for CME-associated and not-associated filaments.

Figure 9 Type distributions for CME-associated and not-associated filaments.

Figure 10 ROC graph showing different SVM topologies with various *d* and *γ* values for validation method 1.

Figure 11 Magnified view of region Z in Figure 10: ROC graph showing the optimum SVM topologies with various *d* and *γ* values for validation method 1. The values (*d*, *γ*) for the optimum topologies are: A(5,8), B(3,6), C(2,8), D(1,1), E(3,10), F(1, 9), G(1, 5), H(7, 8), I(10,7), J(10, 3), K(2,2).

Figure 12 ROC graph showing different SVM topologies with various threshold values for validation method 1.

Figure 13 Magnified view of region Z in Figure 12: ROC graph showing the best SVM topologies with various threshold values for validation method 1. The threshold values for the optimum topologies are: A(0.52), B(0.56), C(0.57), D(0.67), E(0.56), F(0.48), G(0.59), H(0.55), I(0.55), J(0.57), K(0.64).

Figure 14 Comparison among the prediction performances of the current work and all our previous researches on CME prediction. (A) current work, method 1, (B) current work, method 2, (C) Qahwaji *et al.* (2008a), (D) Real and Modest AdaBoost in Qahwaji *et al.* (2008b), (E) Gentle AdaBoost in Qahwaji *et al.* (2008b), (F) SVM in Qahwaji *et al.* (2008c), (G) CCNN in Qahwaji *et al.* (2008c), (H) Al-Omari *et al.* (2008).

Figure 15 ROC graph showing the optimum SVM topologies with various *d* and *γ* values for validation method 2. The values (*d*, *γ*) for the optimum topologies are: A(3,9), B(3,2), C(6,2), D(8,1), E(6,4), F(2,3), G(1,1), H(3,6), I(2,1), J(7,3), K(5,8).

Figure 16 ROC graph showing the best SVM topologies with various threshold values for validation method 2. The threshold values for the optimum topologies are: A(0.58), B(0.44), C(0.64), D(0.50), E(0.47), F(0.48), G(0.44), H(0.56), I(0.72), J(0.44), K(0.47).

Figure 17 A hybrid computer system for CME prediction.

Table 1 Summary of previous research on the associations between CMEs and other solar activities.

Reference	Data	Period	Results related to our work
Munro *et al.* (1979)	75 major Skylab CMEs associated with the solar activity reported at SGD.	1973 to 1974	40% of the CMEs were associated with flares, and 50% of the CMEs were associated with eruptive prominences.
Poland *et al.* (1981)	CMEs were observed from the NRL’s white light coronagraph (Solwind).	1971 to 1974	50% of the CMEs were associated with flares or eruptive prominences.
Webb and Hundhausen (1987)	58 CMEs observed using the HAO Coronagraph/Polarimeter on the SMM satellite.	1980	68% of the CMEs were associated with eruptive prominences and 37% were associated with ~~H-~~Hα flares.
St. Cyr and Webb (1991)	73 CMEs, SMM data.	1984 to 1986	76% of the CMEs were associated with eruptive prominences, 26% were associated with ~~H-~~Hα flares.
St. Cyr *et al.* (1999)	141 CMEs observed using the MK3 K coronameter at MLSO.	1980 to 1989	55% of the CMEs were associated with active regions and 82% were associated with eruptive prominences.
Gilbert *et al.* (2000)	54 ~~H-~~Hα observations obtained from the MLSO.	Feb 1996 to Jun 1998	94% of the eruptive prominences and 46% of the active prominences were associated with CMEs.
Subramanian and Dere	32 CMEs compared with MDI and ~~H-~~Hα images.	Jan 1996 to May 1998	CME associations: 41% with active regions without prominence eruptions,

(2001)			44% with eruptive prominences embedded in active regions, and 15% with eruptive prominences that took place outside active regions.
Hori and Culhane (2002)	50 prominence eruptions near the SM observed using microwave images from the Nobeyama Radioheliograph.	1999 to 2000	92% of the prominence eruptions were associated with CMEs.
Moon *et al*. (2002)	3217 CME events observed using SOHO/LASCO.	1996 to 2000	4% of the CMEs were associated with filaments.
Yang and Wang (2002)	431 filaments compiled from BBSO ~~H-~~Hα images.	Jan 1997 to Jun 1999	30% of the filament disappearances were associated with CMEs.
Gopalswamy *et al*. (2003)	186 prominence eruptions observed using microwave images from the Nobeyama Radioheliograph.	Jan 1996 to Dec 2001	2% of the prominence eruptions were associated with CMEs.
Jing, Yang, and Wang (2003)	79 filaments observed using ~~H-~~Hα or EIT/LASCO.	1999 to 2002	63% of the filaments were associated with CMEs.
Pojoga and Huang (2003)	47 out of 426 disappearing filaments were identified as eruptive filaments.	Jan to Apr 2000	70% of the eruptive filaments were associated with CMEs.
Zhou, Wang, and Cao (2003)	197 front-side halo CMEs observed using SOHO/LASCO.	1997 to 2001	88% of the CMEs were associated with flares and 94% were associated with eruptive filaments.
Jing *et al*. (2004)	106 filament eruptions detected using ~~H-~~Hα images from BBSO.	1999 to 2003	56% of the filament eruptions were associated with CMEs.
Jing (2005)	98 major filament eruption events.	Jan 1999 to Dec 2003	56% of the filaments were associated with CMEs.
Al-Omari *et al*. (2008) and Qahwaji *et al*. (2008a)	All data in SOHO/LASCO CMEs catalogue and NGDC filaments catalogue.	Jan 1996 to Dec 2006	16% of the filaments were associated with CMEs.
Qahwaji *et al*. (2008c)	19164 solar flares and 9297 CMEs.	Jan 1996 to Dec 2004	17.4% of the reported solar flares are CME-associated.

Table 2 Filament types.

Type	Description	Numerical value
SSB	Solar sector boundary	0.10
MDP	Mound prominence	0.15
CRN	Coronal rain	0.20
CAP	Cap prominence	0.25
LPS	Loops	0.30
SPY	Spray	0.35
BSD	Bright surge on disk	0.40
APR	Active prominence	0.45
DSD	Dark surge on disk	0.50
ADF	Active dark filament	0.55
ASR	Active surge region	0.60
AFS	Arch filament system	0.70
BSL	Bright surge on limb	0.75
EPL	Eruptive prominence on limb	0.85
DSF	Disappearing filament	0.90

Table 3 Groups of properties that are used as input nodes in the SVM learning algorithm.

Group	Inputs

4D	Timing, duration, type, extent$_{\text{Disk}}$
4L	Timing, duration, type, extent$_{\text{Limb}}$
3	Timing, duration, type
3aD	Timing, duration, extent$_{\text{Disk}}$
3aL	Timing, duration, extent$_{\text{Limb}}$
3bD	Timing, type, extent$_{\text{Disk}}$
3bL	Timing, type, extent$_{\text{Limb}}$
2	Timing, duration
2a	Timing, type
2bD	Timing, extent$_{\text{Disk}}$
2bL	Timing, extent$_{\text{Limb}}$

Table 4 Averages of performance indicators (Jack-knife technique).

Group	d	γ	TPR	FPR	FNR	TNR	Accuracy	Specificity	Sensitivity	HSS	D_{ROC}	Threshold
4D	5	8	0.85	0.47	0.16	0.53	0.64	0.53	0.85	0.33	0.266	0.52
4L	3	6	0.76	0.53	0.24	0.47	0.53	0.47	0.76	0.21	0.158	0.56
3	2	8	0.65	0.22	0.35	0.78	0.73	0.78	0.65	0.43	0.304	0.57
3aD	1	1	0.37	0.14	0.63	0.86	0.65	0.86	0.37	0.25	0.167	0.67
3aL	3	10	0.75	0.50	0.25	0.50	0.55	0.50	0.75	0.20	0.177	0.56
3bD	1	9	0.61	0.30	0.39	0.70	0.66	0.70	0.61	0.30	0.216	0.48
3bL	1	5	0.59	0.34	0.42	0.66	0.63	0.66	0.59	0.23	0.173	0.59
2	7	8	0.67	0.36	0.33	0.64	0.65	0.64	0.67	0.30	0.219	0.55
2a	10	7	0.62	0.24	0.38	0.76	0.70	0.76	0.62	0.38	0.269	0.55
2bD	10	3	0.46	0.24	0.54	0.76	0.64	0.76	0.46	0.23	0.157	0.57
2bL	2	2	0.32	0.18	0.68	0.82	0.54	0.82	0.32	0.13	0.103	0.64

Table 5 Averages of performance indicators (discarding the extent from inputs).

Group	d	γ	TPR	FPR	FNR	TNR	Accuracy	Specificity	Sensitivity	HSS	D_{ROC}	Threshold
4D′	6	1	0.63	0.24	0.37	0.77	0.71	0.77	0.63	0.39	0.276	0.51
4L′	5	1	0.68	0.41	0.32	0.59	0.58	0.59	0.68	0.25	0.192	0.55
3aD′	1	5	0.45	0.17	0.55	0.83	0.67	0.83	0.45	0.29	0.198	0.49
3aL′	5	1	0.66	0.41	0.34	0.59	0.58	0.59	0.66	0.23	0.180	0.52
3bD′	1	2	0.60	0.28	0.40	0.72	0.67	0.72	0.60	0.32	0.228	0.55
3bL′	2	2	0.77	0.42	0.23	0.58	0.62	0.58	0.77	0.34	0.244	0.49

Table 6 Averages of performance indicators (further validations).

Group	d	γ	TPR	FPR	FNR	TNR	Accuracy	Specificity	Sensitivity	HSS	D_{ROC}	Threshold
4D	3	9	0.32	0.19	0.68	0.81	0.79	0.81	0.32	0.09	0.032	0.58
4L	3	2	0.71	0.36	0.29	0.64	0.64	0.64	0.71	0.25	0.094	0.44
3	6	2	0.64	0.18	0.36	0.82	0.81	0.82	0.64	0.32	0.180	0.64
3aD	1	2	0.12	0.02	0.88	0.98	0.96	0.98	0.12	0.07	0.105	0.66
3aL	6	4	0.29	0.11	0.71	0.90	0.86	0.90	0.29	0.13	0.120	0.47
3bD	2	3	0.53	0.37	0.47	0.63	0.62	0.63	0.53	0.11	0.021	0.48
3bL	1	1	0.71	0.29	0.29	0.71	0.71	0.71	0.71	0.30	0.136	0.44
2	3	6	0.74	0.51	0.26	0.49	0.53	0.49	0.74	0.17	0.042	0.56
2a	2	1	0.62	0.21	0.38	0.79	0.78	0.79	0.62	0.29	0.150	0.72
2bD	8	7	0.94	0.89	0.06	0.11	0.13	0.11	0.94	0.03	0.003	0.47
2bL	5	8	0.36	0.14	0.64	0.86	0.83	0.86	0.36	0.15	0.116	0.47