AUTOMATED MACHINE LEARNING-BASED PREDICTION OF CMES BASED ON FLARES ASSOCIATIONS

R. QAHWAJI, T. COLAK, M. AL-OMARI, and S. IPSON

Department of Electronic Imaging and Media Communications

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

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

Abstract. In this work, machine learning algorithms are applied to explore the relation between significant flares and their associated CMEs. The NGDC flares catalogue and the SOHO/LASCO CMEs catalogue are processed to associate X and M-class flares with CMEs based on timing information. Automated systems are created to process and associate years of flares and CMEs data, which are later arranged in numerical training vectors and fed to machine learning algorithms to extract the embedded knowledge and provide learning rules that can be used for the automated prediction of CMEs. Different properties are extracted from all the associated *(A)* and not-associated *(NA)* flares representing the intensity, flare duration, duration of decline and duration of growth. Cascade Correlation Neural Networks (CCNN) are used in our work. The flare properties are converted to numerical formats that are suitable for CCNN. The CCNN will predict if a certain flare is likely to initiate a CME after input of its properties. Intensive experiments using the Jack-knife techniques are carried out and it is concluded that our system provides an accurate prediction rate of 65.3%. The prediction performance is analysed and recommendation for enhancing the performance are provided.

1. Introduction

The term "space weather" refers to adverse conditions on the Sun, in the solar wind, and in the Earth's magnetosphere, ionosphere, and thermosphere that may affect space-borne or ground-based technological systems and can endanger human health or life Koskinen *et al.* (2001). The importance of space weather is increasing as more human activities take place in space and as we rely more and more on communications and power systems.

The most dramatic solar events affecting the terrestrial environment are solar flares and Coronal Mass Ejections (CMEs) Pick *et al.* (2001). Flares and CMEs are two types of solar eruptions that can spew vast quantities of radiation and charged particles into space Lenz (2004). Earth environment and geomagnetic activity are affected by the ionized solar plasma, also known as the solar wind. Solar wind flows outward from the sun to form the heliosphere and it is affected by solar activity Pick *et al.* (2001) and carries with it the magnetic field of Sun. This interplanetary magnetic field (IMF) creates storms by injecting plasma into the Earth's magnetosphere Yurchyshyn *et al.* (2003); Yevlashin and Maltsev (2003). Geomagnetic storms are correlated with CMEs Wilson and Hildner (1984) and predicting CMEs can be useful to forecasting space weather Webb (2000). Major solar flares can also seriously disrupt the ionosphere and in order to guarantee that humans can work safely and effectively in the space, the forecast for strong solar flares is also important Kurokawa (2002).

Researchers dealing with solar data face many challenges, including the following. Firstly, there is a lack of clear definitions for solar features, which increases the difficulty of designing automated detection and processing systems. Secondly, data

volumes will shortly increase by 1000 to 10,000 times because of the recent space missions (Hinode and STEREO). Extracting useful knowledge from this vast amount of data and trying to establish useful connections between data relating to different time periods is very challenging. Thirdly, large-scale and automated data mining and processing techniques that integrate advanced image processing and machine learning techniques are not fully exploited to find an accurate correlation between the occurrence of solar activities (e.g. flares and CMEs) and solar features observed in various wavelengths.

Despite the recent advances in solar imaging, machine learning and data mining have not been widely applied to solar data. Very recently, several learning algorithms (i.e. neural networks (NN), support vector machines (SVM) and radial basis functions (RBF)) were optimised and then compared for the automated short-term prediction of solar flares Qahwaji and Colak (2007). The machine learning-based system accepts two sets of inputs: The McIntosh classification of sunspot groups and real-time simulation of the solar cycle. Fourteen years of data from the sunspots and flare catalogues of the National Geophysical Data Centre (NGDC) were explored to associate sunspots with their corresponding flares based on their timing and NOAA numbers. Borda *et al.* (2002) described a method for the automatic detection of solar flares using the multi-layer perceptron (MLP) with back-propagation training rule, where a supervised learning technique that required a large number of iterations was used. The classification performance for features extracted from solar flares was compared by Qu *et al.* (2003) using RBF, SVM, and MLPF methods. Each flare is represented using nine features. However, these features provide no information about the position, size and verification of solar flares. Qahwaji and Colak (2006) used NN after image segmentation to verify the regions of interest as solar filaments.

The aim of this paper is to provide a platform for large-scale analysis, association and knowledge extraction for CME and flare data. Data from the publicly available solar flare catalogue, which are provided by the National Geophysical Data Centre (NGDC)[1], is used in our study. NGDC keeps records of data from several observatories around the world and holds one of the most comprehensive publicly available databases for solar features and activities. The CME data are obtained from the SOHO/LASCO CME catalogue, which is generated and maintained by the Center for Solar Physics and Space Weather at the Catholic University of America. This catalogue is developed using the SOHO data in cooperation with the Naval Research Laboratory and the Solar Data Analysis Center (SDAC) at the Goddard Space Flight Center.

This paper is organised as follows: Section 2 explores the association between CMEs and other solar activities or features as reported in previous research. Section 3 describes the design of the Hybrid system and its different components. The practical implementation and evaluation of this system is discussed in Section 4. The concluding remarks and recommendations for future work are presented in Section 5.

2. CMEs and their Associations with Solar Activities and Features

CMEs are bursts of plasma that are ejected from the sun. For years, solar flares were thought to be responsible for major interplanetary (IP) particle events and geomagnetic storms. However, space based chronographs have made us aware of CMEs Tousey (1973). Since then there have been many studies to find out how CMEs are initiated and triggered. The solar flare myth started by Gosling (1995), when it

[1] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/, last access: 2006

was argued that CMEs, not flares, were the critical element for large geomagnetic storms, interplanetary shocks, and major solar energetic particle events. This contradicts the findings of Lin and Hudson (1976) where the flare accelerated particles in big flares are thought to provide the energy for all the activities that followed (i.e., CMEs and large energetic particles events). It is not clear whether there is a cause and effect situation between flares and CMEs and this assumption has driven most of the solar flare myth controversy Cliver and Hudson (2002). A summary of the research on CME associations to other solar features is given below.

An early study was carried out by Munro *et al.* (1979), where 75 major Skylab CMEs from 1973 to 1974 associated with the solar activity reported at Solar Geophysical Data (SGD) are surveyed. It was found that 75% of the CMEs observed were associated with other forms of solar activity, 40% of the CMEs were associated with H-alpha flares, and 50% of the CMEs were associated with eruptive prominences.

Poland *et al.* (1981) used the Naval Research Laboratory's Earth orbiting coronagraph Solwind for observing CMEs. This study was based on white light coronal images from 1971 to 1974. It was concluded that half of the observed CMEs were associated with definite or probable flares or eruptive prominences.

Webb and Hundhausen (1987) compared the CMEs observed in 1980 by the HAO Coronagraph/Polarimeter on the Solar Maximum Mission (SMM) satellite with other forms of solar activity (eruptive prominences (EP), H-alpha flares, soft X-ray events, and metric type II and IV radio bursts). It was found that 66% (38 of 58) of the CMEs were associated with the solar activities under consideration. Out of these CMEs, 68% were found to be associated with eruptive prominences, 37% were associated with H-alpha flares, 76% were associated with X-ray events, and 32% were associated with Radio II, or IV events.

St Cyr and Webb (1991) studied the SMM data from 1984 to 1986 and found that 76% of the CMEs were associated with eruptive prominences, 26% were associated with H-alpha flares and 74% with X-ray events.

Srivastava *et al.* (1997) studied 14 CMEs observed by SMM during the period from March to September 1980 and concluded that strong association existed between CMEs and coronal holes, EPs and current sheets.

St Cyr *et al.* (1999) examined 141 CMEs using Mark III (MK3) K coronameter at Mauna Loa Solar Observatory between 1980 and 1989.They found that 55% of the CMEs were associated with active regions and 82% of the CMEs were associated with the eruption of prominences.

Gilbert *et al.* (2000) examined 54 H-alpha events from February 1996 to June 1998. These prominence events were classified as eruptive prominences and active prominences based on their apparent maximum heights and differences in velocity and acceleration. It is worth mentioning that a variety of prominence classification schemes have been proposed before Gilbert *et al.* (2000), but these classification systems are not entirely compatible with one another. In Gilbert *et al.* (2000), the associations of eruptive prominences and active prominences with CMEs were studied using H-alpha observations that were obtained from Mauna Loa Solar Observatory (MLSO). It was found that 92% of the eruptive prominences and 46% of the active prominences were associated with CMEs.

Subramanian and Dere (2001) studied the sources of 32 CMEs observed between January 1996 and May 1998 and compared them with MDI and several H-alpha images. It was found that 41% of the CMEs were associated with active regions without prominence eruptions, 44% were associated with eruptive prominences

embedded in active regions, and 15% were associated with eruptive prominences that have taken place outside active regions.

Hori and Culhane (2002) used microwave images from Nobeyama Radioheliograph, to examine 50 prominence eruptions near solar maximum between 1999 to 2000, and showed that 92% of the prominence eruptions were associated with CMEs.

Zhou *et al.* (2003) examined the correlation between Halo CMEs and solar surface activity observed by SOHO/LASCO from 1997 to 2001. They concluded that 88% of halo CMEs were associated with flares and more than 94% were associated with eruptive prominences/filaments, while 79% of the CMEs were initiated from active regions.

Moon *et al.* (2002) analysed 3217 CME events observed by SOHO/LASCO from 1996 to 2000 and made a statistical study on their association with solar flares using GOES X-ray images and eruptive filaments using H-alpha images from BBSO. They found that the flares that were associated with CMEs had larger velocities. Macqueen and Fisher (1983) and Sheeley *et al.* (1999)drew similar conclusions that if no flare or only a weak flare occurs, then we would have the slowly-accelerating eruptive filament events but if a flare occurs, then an additional acceleration process might act on the CME. The key observation is the acceleration profile of the CME (or filament) during the flare.

Gopalswamy *et al.* (2003) classified the prominence eruptions as radial and traverse depending on the direction of their movement (radial or horizontal). The associations with CMEs were investigated as well. Microwave images of 186 prominence eruptions from Nobeyama radioheliograph from 1 January 1996 to 31 December 2001, covering the minimum and maximum periods of the current solar cycle 23 were used. It was found that 152 images (82%) of the prominence eruptions were dominantly radial events while only 34 images (18%) were traverse events and 134 images (72%) of the prominence eruptions were found to be clearly associated with CMEs. They also found that 83% of the radial events were associated with CMEs.

Pojoga and Huang (2003) studied the sudden disappearances (commonly called Disparition Brusque) of prominences/filaments identified from the H-alpha images of Praire View Solar Observatory and the spectroheliograms of Meudon Observatory for the period from January to April 2000 and their correlation with CMEs. According to this study 70% of the eruptive filaments were associated with CMEs, while the correlation was weaker for the quasi-eruptive and vanishing filaments. In this study, they used the term “vanishing” when referring to the thermal disappearances of prominences/filaments.

Mouradian *et al.* (1995) analysed the DB of quiescent filaments/prominences in two classes; dynamic and thermal disappearances. They considered the dynamic DB to consist of an expansion and ejection of prominence plasma into the corona due to changes in the underlying magnetic field structure, like the emergence of new magnetic flux. On the other hand, the thermal DB consisted of the disappearance of prominences in H-alpha line due to an energy increase. This study showed that dynamic DBs were associated with CMEs, whereas thermal DBs were just local disturbances at the lower corona.

Jing *et al.* (2004) performed a statistical study of 106 filament eruptions detected using H-alpha images from BBSO from 1999 to 2003 and their relations to flares and CMEs. According to their study 56% of the filament eruptions were associated with CMEs. They also classified filament eruptions as active region filament eruptions and quiescent filament eruptions and found that active region filament eruptions had higher flare association (95%) compared to quiescent filament eruptions (27%). They

found that quiescent filament eruptions were mostly accompanied by CMEs rather than flares. They also suggested that the emergence of new flux played an important role in destabilising filaments.

Not all researchers agree that strong relations exist between CMEs and filament/prominence eruptions. Yang and Wang (2001)made a statistical study of 431 filament/prominence disappearances compiled from BBSO H-alpha images observed between January 1997 and June 1999 and found that a low association with CMEs existed (only 30%). However, they stated that they didn't make a distinction between thermal filament disappearances and filament eruptions. In addition, filament disappearances on disk might be associated with very weak halo CMEs which were difficult to detect. On the other hand, there are also many case studies on the solar origin of CMEs including Zhang and Wang (2001); Webb *et al.* (1998), which found associations between large CMEs and filament eruptions.

In Green *et al.* (2003) flares were examined in nine active regions with CME signatures. It was indicated that the energy released by flaring from the magnetic field of an active region was greater anterior to the CME launch than after. The research of Zhang and Wang (2001) measured the CMEs initial evolution in the low corona and then explored the possible causes of CME initiation and acceleration in connection with flares. The kinematical evolution of CMEs is described in a three-phase scenario: the initiation phase, the impulsive acceleration phase, and the propagation phase.

The changes associated with the magnetic topology for the X1.2 flare that occurred on 30 September 2000 and was not associated with a CME were studied in Green *et al.* (2003). It was noted that the flare resulted from the interaction of two pre-existing loops low in the corona which produced a confined flare.

In Andrews (2003), 311 M and X-class flares, which occurred during the years 1996 to 1999, were investigated to find their associated CME candidates. The SOHO/LASCO CME data were used in this study. Online catalogues were used to search for CME candidates for the 229 flares with good LASCO data coverage. It was found that about 40% of the M-class flares do not have associated CMEs.

In Akiyama *et al.* (2006) the CME association rate for two flares, which were produced by two active regions, was examined. Active region 10039 produced three X- and eight M-class flares and the CME-Flare association rate was found to be 72%. The CMEs from this active region had an average speed of 1195 km/s speed and an average width of 246°. On the other hand, active region 10044 produced 9 M-class flares, the association rate was found to be 13%, and CMEs from this region had an average speed of 282 km/s speed and an average width of 12°.

A statistical analysis of the latitudinal locations for the flares in northern and southern hemispheres for the period of 1986 to 2003 was conducted in Shrivastava and Singh (2005). It was found that flares with associated CMEs are equally distributed in the northern and southern hemispheres. It is noted that flares associated with northern hemisphere CMEs are more likely to produce a rapid decrease in the observed galactic cosmic ray (high-energy charged particles) intensity following a CME which is known as Forbush decrease. It occurs due to the magnetic field of the plasma solar wind sweeping some of the galactic cosmic rays away from Earth.

There is also some research on the intensity of the solar flares and CMEs. Yashiro *et al.* (2005) examined the CME visibility (detection efficiency) for 1301 X-ray flare events above C3 level (49 X-class, 610 M-class, and 642 C-class flares) from 1996 to 2001. It is assumed that all CMEs associated with limb flares are detectable by LASCO. Based on a statistical study of the properties of the flare-associated CMEs and a comparison with flare size and longitude it was found that the CME association

rate increased with the flare size from 20% for C-class flares to 100% for huge X-class flares. It was also concluded that CMEs associated with disk C-class flares were slower and narrower than those of CMEs associated with X-class flares.
A discussion of the associations of CMEs with flare properties is presented in Yashiro *et al.* (2006). Properties such as peak X-ray intensity, total X-ray intensity, and the decay time for 1540 X-ray flares (M-class and above, including 50 huge flares above X1.8) were analyzed. It was found that CMEs associated with flares above X1.8 have CME association rate of 98% compared with only 40% for CMEs associated with flares between M1.0 and M1.7. Also it was concluded that a definite association between CMEs and flares exists if the decay time of the flare exceeds 90 min.

3. Designing the Computer Platform for CMEs Prediction

In the survey of section 2, it is clear that there has been no large-scale processing and analysis for all the available records of CMEs and flares to determine their association. Most of the available studies are carried out on a few years of data or on limited cases. In this work, we present a computer platform that analyses all the available years of data related to flare and CME catalogues to extract learning rules and then provide automated prediction for CMEs. Several different stages are involved in this system, as shown in Figure 1 and are explained in the following sections.

3.1 ASSOCIATING FLARES AND CMES

A C++ platform is created to automatically associate CMEs in the SOHO/LASCO CMEs catalogue with flares in the NGDC X-ray flares catalogue. The association is determined based on their timing information; the date and time for every CME is compared with date and time for every flare.

Two criteria are used for comparison:

- If there is not a CME recorded "α" minutes before or after a flare reach its peak time, then this flare is marked as not-associated (*NA*) otherwise it is marked as possibly-associated (*PA*).
- If there is a CME recorded "β" minutes after a *PA* flare reaches its peak then this flare is marked as an associated *(A)* flare, other wise it is marked as *NA*.

After finding all the associations, a numerical dataset is created for the machine learning algorithms using *not-associated* and *associated* flares.

3.2 CREATING THE ASSOCIATED NUMERICAL DATA SET

In this work, we have processed all the CME and flare data for the period from January 1996 until the end of December 2004. Our software has analyzed the data relating to 9297 CMEs and 19164 flares, which occurred during this period.
To determine the *NA* flares the value of α was made equal to 150 minutes in all our experiments. It is easier to determine if a CME is not associated with any flares rather than determine the level of association between every CME with flares based on timing information. To explore the different levels of associations we have applied our association algorithm with different values of β, as shown in Table 1. As expected, more CMEs are associated with flares as the value of β increases. The rate of increase in the number of associations is maximum, when β increases to 60 from 30 minutes. The rate of increase is equal to 85%, 33% and 23% when β increases to 60 from 30, to 90 from 60 and to 120 from 90, respectively. Since the increase in the association rate drops from 85% to 33% after the 60 minutes difference, this makes β = 60 suitable for our experiments.

In this paper, CMEs are associated with significant flares (i.e., X and M- class flares) only. X and M-class flares can have significant impacts on our life on Earth. In our previous work Qahwaji and Colak (2007) an automated machine-learning system that can provide short-term prediction for the occurrences of these significant flares is introduced. Our long-term goal is to determine the level of associations between CMEs and flares using machine learning so that a hybrid system that integrates both systems can be designed. Associating CMEs with significant flares seems to be supported by the findings of Yashiro *et al.* (2005), where it was found that all CMEs associated with X-class flares are detected by LASCO, while almost half the CMEs associated with C flares are invisible. They also concluded that the CME association rate increases with the increase of the x-ray brightness for flares starting from 20% for C-class flares (between C3 and C9 levels) to 100% for huge flares (above X3 level). In addition, they found that the CMEs that are faster (median 1556 km/s) and wider (median 244°) are associated with X-class flares compared to the CMEs associated with disk C-class flares (432 km/s, 68°).

By applying our association algorithm we created an associated data set. This set consists of 985 flares with 581 *A* flares and 404 *NA* flares with $\alpha = 150$ minutes and $\beta = 60$ minutes. Because machine learning algorithms deal mainly with numbers, it is essential that appropriate numerical representations for the *A* and *NA* flares is proposed and implemented. We can extract properties such as intensity, starting time, peak and ending time of the flares from the NGDC flares catalogue. However, we were hoping to include additional properties for flares location. Unfortunately a large number of the associated flares don't have their location information in the NGDC catalogues. Hence, we decided to use the properties shown in Table II. Numerical representations are needed for these properties which are used later to construct different input parameters for the training and testing stages of the machine learning system. As we are not sure which properties are more important for machine learning and for the prediction of CMEs we decided to carry out extensive experiments in order to determine the significance of each property for our application.

4. Practical Implementation and Results

After creating the associated data set, the training and testing experiments for the machine learning algorithm was begun. These experiments and the prediction performance are explained below.

4.1 TABOUT THE LEARNING ALGORITHMS AND TECHNIQUES

For our study, we have compared the performance of Cascade Correlation Neural Network (CCNN) and Support Vector Machines (SVM) which have proven to be a very effective learning algorithm for similar applications Qahwaji and Colak (2007). More information on the theory and implementation of these learning algorithms are provided in Qahwaji and Colak (2007). All the machine learning/training and testing experiments are carried out with the aid of the Jack-knife technique Fukunaga (1990). This technique is usually implemented to provide a correct statistical evaluation for the performance of the classifier when implemented on a limited number of samples. This 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 decide which samples are used for the training of the classifier and which are kept for testing it. The classification error depends mainly on the training and testing samples. For a finite number of samples, the error counting procedure can be used to estimate the performance of the Fukunaga (1990). In each experiment, 80% of the samples were

randomly selected and used for training while the remaining 20% were used for testing. For every reported performance value to follow, this value is obtained by carrying out 10 Jack-Knife experiments and finding their average value.

4.1 Optimising the learning algorithms

The prediction performance of CCNN and SVM are compared to determine the machine learning algorithm that is more suitable for our application. However, these learning algorithms must be optimised before the actual comparison can take place. Learning algorithms are optimised to ensure that their best performances are achieved. In order to find the best parameters and/or topologies for the three learning algorithms initial training and testing experiments are applied using the Jack-Knife technique as explained previously. The results of these experiments are used to determine the optimum parameters and topology for every machine learning algorithm before it can be compared with the other learning algorithm. The optimisation process for the learning algorithms used in this work is described below:

- For every learning algorithm apply the learning experiments with one, two and three inputs.
- For each learning experiment determine the best topology using ROC analysis.
- Determine the optimum classification threshold for the optimum topology using ROC curves.
-

Finding the optimum topology for CCNN is reached by determining the number of input features and means determing the process for CCNN consists of finding the optimum topology, while the optimisation process for SVM consists of finding the correct values of gamma and degree for the Anova kernel. ROC analysis is applied in all these experiments.

4.1.2 Optimising the CCNN

In Qahwaji and Colak (2006b), it was proven that CCNN provides the optimum neural network performance for processing solar data in catalogues. However, many hidden nodes and just one hidden layer were used for training the network in Qahwaji and Colak (2006b). To simplify the topology of CCNN more experiments with two hidden layers are concluded in this work by changing the number of hidden nodes in each layer from one to ten. At the end we managed to compare 100 different CCNN topologies based on the best CFP and CFTP. As can be seen from Figures 3 and 4, a CCNN with six hidden nodes in the first layer and four hidden nodes in the second layer gives the best results for CFP and CFTP.

The CCNN that we have used consists of input, hidden and output layers. The output layer consists of one output node which has a numerical value of 0.9 if a CME is predicted to occur and 0.1 if not. To find the optimum topology that will provide the best learning performance we have carried out large number of experiments in a manner similar to Qahwaji and Colak (2007).

The number of input parameters/nodes and the number of hidden nodes in each experiment were changed to find the best parameters and topologies for this learning algorithm. The number of input parameters/nodes was changed from one to four and the number of hidden nodes was changed from one to ten. For each topology we applied Jack-knife technique ten times and recorded the average of the testing results for correct prediction rate of CMEs.

We started our experiments with one input node, which represented the numerical representation for property *A.* After changing number of hidden nodes from one to ten by applying the Jack-knife technique ten times for each new hidden node, a second input node representing property *B* was added and the experiments were repeated as explained before. The same procedure was repeated for property *C* and then for property *D*. Using this method we compared 40 (4 x 10) different neural network topologies by testing each topology ten times and recording the average correct CME prediction rate.

We used the MATLAB neural network toolkit for our experiments. We applied 788 *associated* and *not-associated* flares for training, which are the 80% of the total number of associated cases. We have also used 197 *associated* and *not-associated* flares for testing, which are the 20% of the total number of associated cases. Figure 2 shows the average of correct CME prediction rates for each neural network topology.

4.2. COMPARING THE PREDICTION PERFORMANCES

The results show that there is an increase in the prediction rate every time a new input parameter is added except for the case when the incline duration of flare (property *D*) is added. This shows that the time needed for a flare to reach its peak intensity is not very important in terms of CME predictions using machine learning. Also, we found that the decline duration for the flare (property *C*), is more important for CMEs prediction than the total flare duration (property *B*). This means that decline duration of the flare is very important for determining the probability of CME occurrence and this coincides with the findings of Yashiro *et al.* (2006), as explained in our survey section.

We have concluded that the best topology for CME predictions using flares is to use a CCNN with three input nodes representing the intensity of flare, flare duration, and decline duration. The optimum topology consists also of ten hidden nodes. Using the optimum topology we have conducted ten experiments where for each the training set contains 80% randomly selected cases. The prediction rates are obtained by testing the learning system with the remaining 20% of the associated cases. This has lead to an average CMEs prediction rate of 65.3%.

5. Conclusions and Future Research

In this paper, a machine-learning based system that analyses years of flares and CMEs data is introduced. This system analyses all data records in the NGDC flares catalogues and the SOHO/LASCO CMEs catalogue and applies our association algorithm to associate flares with their corresponding CMEs based on timing information. In our work, we have used the CCNN because of its efficient knowledge extraction and generalisation performance for the processing of solar data Qahwaji and Colak (2007). To determine the optimum CCNN topology that delivers the best learning and then prediction performance, many experiments are carried out by changing the number of input and hidden nodes. It was found that a CCNN with three input nodes, ten hidden nodes and one output nodes provides the best prediction performance.

We have investigated all the reported flares and CMEs between 01 January 1992 and 31 December 2004. Our software has managed to associate 581 M and X soft X-ray flares with their corresponding CME groups and managed to highlight another 404 significant flares as being not associated with any CMEs. These associations are for α = 150 minutes and β = 60 minutes. After carrying out many experiments using the Jack-knife technique, an average CMEs prediction rate of 65.3% was achieved. In its

current version, our system can provide one hour prediction in advance, by analysing the latest flares data.

We believe that our work is the first to introduce a fully automated computer platform that could verify the association between significant flares and CMEs using machine learning. This work is a first step towards constructing a fully automated and web-compliant platform that would provide short-term prediction for the possible eruptions of CMEs. In this paper, we managed to extract the experts' knowledge which is embedded in the CMEs and flares catalogues and managed to represent this knowledge using association and learning algorithms. However, our work is far from complete because of the following:

1. We have managed to associate only a small percentage of CMEs with significant flares (M and X-class flares). However, the largest rate of association is for CMEs associated with C-class flares, as shown in Table 1. For our future work we will investigate the association for the C and B-class flares as well.
2. The prediction performance is not as high as we would like it to be. We believe that this is caused by:

a. From the survey section provided here, it is obvious that CMEs can be associated with either flares or erupting filaments/prominence. In this study, CMEs were associated with flares only and erupting filaments/prominence are not considered. To enhance the accuracy our predictions CMEs that are associated with eruptive filaments have to be considered. For example, on 21 March 1999 a filament erupted from the southern boundary of NOAA AR8494. The filament erupted between 12:35 and 14:30 UT. Its associated CME first appeared in the field of view of the LASCO C2 at 15:54UT, and later in the LASCO C3 at 17:42UT. This CME is not associated with any significant X-ray flare or H-alpha flare, as studied in Yun-Chun *et al.* (2006). This coincides with the results provided by our association algorithm, which highlights this CME as a not associated CME.

b. The association between flares and CMEs in our work is carried out based on time analysis only. As explained in Yashiro *et al.* (2006) this may lead to false associations. There is a small difference in the visibility of CMEs between front side and backside CMEs, which makes it very hard to distinguish them using coronagraph observations only Yashiro *et al.* (2006). To overcome this situation we need to confirm that the CME originates from the front side by checking the lower corona images obtained by EIT and SXT. This will be investigated in our future work.

Acknowledgment

This work is supported by an EPSRC Grant (GR/T17588/01), which is entitled "Image Processing and Machine Learning Techniques for Short-Term Prediction of Solar Activity".

References

Akiyama, S., Gopalswamy, N. & Yashiro, S. 2006, *36th COSPAR Scientific Assembly* (Beijing, China), p. 556.

Andrews, M.D. 2003, Sol. Phys. 218, 261.

Borda, R.A.F., Mininni, P.D., Mandrini, C.H., Gomez, D.O., Bauer, O.H. & Rovira, M.G. 2002, Sol. Phys. 206, 347.

Cliver, E.W. & Hudson, H.S. 2002, Journal of Atmospheric and Solar-Terrestrial Physics 64, 231.

Fukunaga, K. 1990, *Introduction to Statistical Pattern Recognition," Academic Press, New York, 1990.*

(New York: Academic Press).

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.

Gosling, J.T. 1995, J. Geophys. Res-Space Phys. 100, 7921.

Green, L.M., Lopez Fuentes, M.C., Mandrini, C.H., van Driel-Gesztelyi, L. & Demoulin, P. 2003, Advances in Space Research 32, 1959.

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

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

Koskinen, H., Tanskanen, E., Pirjola, R., Pulkkinen, A., Dyer, C., Rodgers, D. & Cannon, P. 2001, *ESA Space Weather Programme Feasibility Studies* (FMI, QinetiQ, RAL Consortium).

Kurokawa, H. 2002, Journal of the Communications Research Laboratory. Special issue on Space Weather Forecast Study on Space Weather and its Hazards 49.

Lenz, D. 2004, The Industrial Physicist 9, 18.

Lin, R.P. & Hudson, H.S. 1976, Sol. Phys. 50, 153.

Macqueen, R.M. & Fisher, R.R. 1983, Sol. Phys. 89, 89.

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

Mouradian, Z., Soruescaut, I. & Pojoga, S. 1995, Sol. Phys. 158, 269.

Munro, R.H., Gosling, J.T., Hildner, E., Macqueen, R.M., Poland, A.I. & Ross, C.L. 1979, Sol. 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, Advances in Space Research 32, 2641.

Poland, A.I., Howard, R.A., Koomen, M.J., Michels, D.J. & Sheeley, N.R., Jr. 1981, Sol. Phys. 69, 169.

Qahwaji, R. & Colak, T. 2006, The International Journal of Computers and Their Applications 13, 9.

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

Qu, M., Shih, F.Y., Jing, J. & Wang, H.M. 2003, Sol. Phys. 217, 157.

Sheeley, N.R., Walters, J.H., Wang, Y.M. & Howard, R.A. 1999, J. Geophys. Res-Space Phys. 104, 24739.

Shrivastava, P.K. & Singh, N. 2005, Chin. J. Astron. Astrophys. 5, 198–202.

Srivastava, N., Gonzalez, W.D. & Sawant, H.S. 1997, Advances in Space Research 20, 2355.

St Cyr, O.C., Burkepile, J.T., Hundhausen, A.J. & Lecinski, A.R. 1999, J. Geophys. Res-Space Phys. 104, 12493.

St Cyr, O.C. & Webb, D.F. 1991, Sol. Phys. 136, 379.

Subramanian, P. & Dere, K.P. 2001, Astrophys. J. 561, 372.

Tousey, R. 1973, Adv. Space Res. 13, 713.

Webb, D.F. 2000, Journal of Atmospheric and Solar-Terrestrial Physics 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, Sol. Phys. 108, 383.

Wilson, R.M. & Hildner, E. 1984, Sol. Phys. 91, 169.

Yang, G. & Wang, H. 2001, *Solar-Terrestrial Magnetic Activity and Space Environment, Proceedings of the COSPAR Colloquium held in the NAOC*. Wang, H. & Xu, R. (eds.) (Beijing, China), vol. 14, p. 113.

Yashiro, S., Gopalswamy, N., Akiyama, S. & Howard, R.A. 2006, *36th COSPAR Scientific Assembly* (Beijing), p. 1778.

Yashiro, S., Gopalswamy, N., Akiyama, S., Michalek, G. & Howard, R.A. 2005, Journal of Geophysical Research 110.

Yevlashin, L.S. & Maltsev, Y.P. 2003, Geomagn. Aeron. 43, 269.

Yun-Chun, J., Le-Ping, L. & Li-Heng, Y. 2006, Chin. J. Astron. Astrophys. 6, 345.

Yurchyshyn, V., Wang, H. & Abramenko, V. 2003, Advances in Space Research 32, 1965.

Zhang, J. & Wang, J.X. 2001, Astrophys. J. 554, 474.

Zhou, G.P., Wang, J.X. & Cao, Z.L. 2003, Astron. Astrophys. 397, 1057.

Figure 1: The hybrid prediction computer system.
Figure 2: The prediction performance for CCNN.

Table 1: The levels of associations based on the value of β

Flares	X	M	C	B	Total
NA	15	389	5554	3355	9313
PA(α=150)	89	926	6770	2066	9851
Total	**104**	**1315**	**12324**	**5421**	**19164**
A(β=30)	57	318	1181	246	1802
A(β=60)	71	510	2229	526	3336
A(β=90)	77	592	3016	764	4449
A(β=120)	78	654	3757	1018	5507

Table 2: Description of each property that is used as input node in CCNN.

	Name	Description
A	***Intensity***	The normalized numerical value of intensity of the flare (Intensity × 1000).
B	***Flare Duration***	The normalized numerical value of the time difference in minutes between the ending and the starting time of the flare (Difference/120).
C	***Decline Duration***	The normalized numerical value of the time difference in minutes between the ending and the peak time of the flare (Difference/120).
D	***Incline Duration***	The normalized numerical value of the time difference in minutes between the peak and the starting time of the flare (Difference/120).