

AUTOMATIC SHORT-TERM SOLAR FLARE PREDICTION USING MACHINE LEARNING AND SUNSPOT ASSOCIATIONS

R. QAHWAJI and T. COLAK

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)

Abstract. In this paper, a machine learning-based system that could provide automated short-term solar flares prediction is presented. This system accepts two sets of inputs: McIntosh classification of sunspot groups and solar cycle data. In order to establish a correlation between solar flares and sunspot groups, the system explores the publicly available solar catalogues from the National Geophysical Data Centre (NGDC) to associate sunspots with their corresponding flares based on their timing and NOAA numbers. The McIntosh classification for every relevant sunspot is extracted and converted to a numerical format that is suitable for machine learning algorithms. Using this system we aim to predict if a certain sunspot class at a certain time is likely to produce a significant flare within six hours time and whether this flare is going to be an X or M flare. Machine learning algorithms such as Cascade-Correlation Neural Networks (CCNN), Support Vector Machines (SVM) and Radial Basis Function Networks (RBFN), are optimised and then compared to determine the learning algorithm that would provide the best prediction performance. It is concluded that SVM provides the best performance for predicting if a McIntosh classified sunspot group is going to flare or not but CCNN is more capable of predicting the class of the flare to erupt. A hybrid system that combines SVM and CCNN is suggested for future use.

1. Introduction

Solar activity is the driver of space weather. Thus it is important to be able to predict the violent eruptions such as coronal mass ejections (CME) and solar flares (Koskinen *et al.*, 1999). An efficient space weather forecasting system should analyse the recent solar data and provide reliable predictions of forthcoming solar activities and their possible effects on Earth. Space weather is defined by the U.S. National Space Weather Program (NSWP) as “conditions on the Sun and in the solar wind, magnetosphere, ionosphere, and thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life or health”. The importance of understanding space weather is increasing for two reasons. First the way solar activity affects life on Earth. Second we rely more and more on communications and power systems, both of which are vulnerable to space weather.

Space weather forecasting is an emerging science that is facing many challenges. Some of these challenges are the automated definition, classification and representation of solar features and the establishing of an accurate correlation between the occurrence of solar activities (e.g., solar flares and CME) and solar features (e.g., sunspots, filaments and active regions) observed in various wavelengths.

In order to predict solar activities we need to use real-time, high-quality data and data processing techniques (Wang *et al.*, 2003). Solar flares research has shown that flares are mostly related to sunspots and active regions (Liu *et al.*, 2005; Zirin and Liggett, 1987; Shi and Wang, 1994). In this paper we aim to design an automated system that uses machine learning algorithms to provide efficient short-term prediction of significant flares. The training of machine learning algorithms is based on associating sunspots with their corresponding flares and on the numerical simulation of the solar

cycle. In the near future we intend to extend the work reported in this paper and combine it with automated image processing algorithms that will analyse solar images in real-time mode. Solar features will be extracted and classified in real-time and then fed to the machine learning algorithms reported here to provide real-time automated short-term prediction of solar flares.

There have been noticeable developments recently in solar imaging and the automated detection of various solar features, by: Gao *et al.* (2002), Turmon *et al.* (2002), Shih and Kowalski (2003), Benkhalil *et al.* (2003), Lefebvre and Rozelot (2004) and Qahwaji and Colak (2006a). Despite the recent advances in solar imaging, machine learning and data mining have not been widely applied to solar data. However, there is no machine learning algorithm that is known to provide the “best” learning performance especially in the solar domain. One of the early attempts to predict solar activity was reported in Calvo *et al.* (1995). However in Calvo *et al.* (1995) solar activity was represented in terms of the Wolf number, not in terms of flares or CMEs. A recent survey describing the imaging techniques used for solar feature recognition is presented in Zharkova *et al.* (2005). 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 and used supervised learning technique that required large number of iterations. Qu *et al.* (2003) compared the classification performance for features extracted from solar flares using Radial Basis Function Networks (RBFN), Support Vector Machines (SVM) and MLP methods. Each flare is represented using nine features. However, these features provide no information about the position, size and classification of solar flares. Neural Networks (NNs) were used in Zharkova and Schetin (2003) for filament recognition in solar images and in Qahwaji and Colak (2006a) NNs were used after image segmentation to verify the regions of interest, which were solar filaments.

There is also some work reported on the correlation between sunspot classifications and flare occurrences and prediction. NOAA Space Environment Laboratory designed an expert system called THEO (McIntosh, 1990) that uses flare rates of sunspot groups classified by the McIntosh scheme, dynamical properties of sunspot growth, rotation, shear and previous flare activity, for assigning a 24 hour probabilities for X-ray flares. The system was subjective because the forecast decision rules were incorporated by a human expert and had not been evaluated statistically. In Sammis *et al.* (2000), eight years of active region observations from the US Air Force/Mount Wilson data set, supplied by the NOAA World Data Centre were studied to explore the relation between δ sunspots and large flares. They confirmed K unzel (1960)’s findings that regions classified as $\beta\gamma\delta$ produce more and larger flares than other regions of comparable size. In particular, they concluded that each region larger than 1000 millionths of hemisphere and classified as $\beta\gamma\delta$ had nearly 40% probability of producing X1 class flares or even bigger flares. Bornmann and Shaw (1994) used multiple linear regression analysis to derive the effective solar flare contributions of each McIntosh classification parameters. They concluded that the original seventeen McIntosh Classification parameters can be reduced to ten and still reproduce the observed flare rates. Gallagher *et al.* (2002) provided a method for finding the flaring probability of an active region in a given day using Poisson statistics and historical averages of flare numbers for McIntosh classifications. Wheatland (2004) introduced a Bayesian approach to flare prediction. This approach can be used to improve or refine the initial prediction of big flares during a subsequent period of time using the flaring record of an active region together with phenomenological rules of flare statistics. To construct the flaring records for an active region, all the flares (whether significant or

not) are considered. This method was tested on limited dataset (i.e., ten flaring days) but the author made interesting points on how it can be adapted to handle real-time data. More recently, Ternullo *et al.* (2006) presented a statistical analysis carried out by correlating sunspot group data collected at the INAF-Catania Astrophysical Observatory with data of M- and X- class flares obtained by the GOES-8 satellite in the soft X-ray range during the period from January 1996 till June 2003. It was found that the flaring probability of sunspots increases slightly with the sunspot age as the flaring sunspots have longer life times compared to non-flaring ones. These previous works demonstrate that there is a clear correlation between sunspot classification and the flaring probability. However, this relation hasn't been verified on large-scale data using automated machine learning algorithms. In this work, we aim to design a fully automated system that can analyse all the reported flares and sunspots for the last fourteen years to explore the degree of correlation and provide an accurate prediction using time-series data. The work reported in this paper is a first step toward constructing a real-time, automated, web compliant system that will extract the latest solar images in real-time and provide automated prediction for the occurrence of big flares within six hours time period.

This paper is organised as follows: the solar data used in this paper are described in Section 2. Machine learning algorithms are discussed in Section 3. The implementation of the system and experimental results are reported in Section 4. Finally, the concluding remarks are given in Section 5.

2. Data

In this work, we have used data from the publicly available sunspot group catalogue and the solar flare catalogue. Both catalogues are provided by the National Geophysical Data Centre (NGDC)¹. NGDC keeps record of data from several observatories around the world and holds one of the most comprehensive publicly available databases for solar features and activities.

Flares are classified according to their X-ray brightness in the wavelength range from one to eight Angstroms. *C*, *M*, and *X* class flares can affect earth. *C*-class flares are moderate flares with few noticeable consequences on Earth (i.e., minor geomagnetic storms). *M*-class flares are large; they generally cause brief radio blackouts that affect Earth's polar regions by causing major geomagnetic storms. *X*-class flares can trigger planet-wide radio blackouts and long-lasting radiation storms. The NGDC flare catalogue (Figure 1) provides information about dates, starting and ending times for flare eruptions, location, X-ray classification, and the NOAA number for the active regions that are associated with the detected flares.

The NGDC sunspot catalogue (Figure 1) holds records of many solar observatories around the world that have been tracking sunspot regions and supplying their date, time, location, physical properties, and classification. Two classification systems exist for sunspots: McIntosh and Mt. Wilson. McIntosh classification depends on the size, shape and spot density of sunspots, while the Mt. Wilson classification (Hale *et al.*, 1919) is based on the distribution of magnetic polarities within spot groups (Greatrix and Curtis, 1973). In this study we used the McIntosh classification which is the standard for the international exchange of solar geophysical data. It is a modified version of the Zurich classification system developed by Waldmeier. The general form of the McIntosh classification is Zpc where, Z is the modified Zurich class, p is the type

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

of spot, and c is the degree of compactness (distribution of sunspots) in the interior of the group.

These two catalogues are analyzed by a C++ computer program that we have created, in order to associate sunspots with their corresponding flares using their timing and location information. This computer program provides training and testing data sets for machine learning algorithms using these associations. More information is provided about the program and data sets in section 4.

3. Machine Learning Algorithms

In this work, Cascade-Correlation Neural Networks (CCNN), Support Vector Machines (SVM) and Radial Basis Function Networks (RBFN) are used and compared for flare prediction. CCNN and RBFN are used because of their efficient performance in applications involving classification and time-series prediction (Frank *et al.*, 1997). In our previous work Qahwaji and Colak (2006b) the performance of several NN topologies and learning algorithms was compared and it was concluded that CCNN provides better association between solar flares and sunspot classes. It is one of the recent trends in machine learning to compare the performance of SVM and neural networks. The work reported in Acir and Guzelis (2004), Pal and Mather (2004), Huang *et al.* (2004), and Distante *et al.* (2003) though not related to the prediction of solar activities, has shown that SVM have outperformed neural networks in the classification tasks for different applications. Similar performance for SVM were reported for flares detection in Qu *et al.* (2003), where statistical features were extracted from solar images and were trained using MLP, RBFN and SVM for automatic flares detection. Many existing references, such as Qu *et al.* (2003), provide detailed description for SVM and RBF. However, the same is not true for CCNN, so we have decided to provide more details for CCNN.

3.1. CASCADE-CORRELATION NEURAL NETWORKS (CCNN)

The training of Backpropagation neural networks is considered to be slow because of the *step-size problem* and the *moving target problem* (Fahlmann and Lebiere, 1989). To overcome these problems cascade neural networks were developed. The topology of these networks is not fixed. The supervised training begins with a minimal network topology. New hidden nodes are added gradually to create a multi-layer structure. The new hidden nodes are added to maximize the magnitude of the correlation between the new node's output and the residual error signal we are trying to eliminate. The weights of every new hidden node are fixed and never changed later, hence making it a permanent feature-detector in the network. This feature detector can then produce outputs or create other more complex feature detectors (Fahlmann and Lebiere, 1989).

In CCNN, the number of input nodes is determined based on the input features, while the number of output nodes is determined based on the number of different output classes. The learning of CCNN starts with no hidden nodes. The direct input-output connections are trained using the entire training set with the aid of the backpropagation learning algorithm. Hidden nodes are then added gradually and every new node is connected to every input node and to every pre-existing hidden node. Training is carried out using the training vector and the weights of the new hidden nodes are adjusted after each pass (Fahlmann and Lebiere, 1989).

However, a major problem facing these networks is over-fitting the training data, especially when dealing with real-world problems (Smieja, 1993). Over-fitting usually

occurs if the training data are characterized by many irrelevant and noisy features (Schetinin, 2003). On the other hand, the Cascade-Correlation architecture has several advantages. Firstly, it learns very quickly, at least 10 times faster than Back-propagation algorithms (Shet *et al.*, 2005). Secondly, the network determines its own size and topology and it retains the structures it has built even if the training set changes (Fahlmann and Lebiere, 1989). Thirdly, it requires no back-propagation of error signals through the connections of the network (Fahlmann and Lebiere, 1989). Finally, this structure is useful for incremental learning in which new information is added to the already trained network (Shet *et al.*, 2005).

3.2. SUPPORT VECTOR MACHINE (SVM)

SVMs are becoming powerful tools for solving a variety of learning and function estimation problems. Compared to neural networks, SVMs have a firm statistical foundation and are guaranteed to converge to a global minimum during training. They are also considered to have better generalisation capabilities than neural networks. SVMs have been shown to outperform neural networks in a number of applications (Qu *et al.*, 2003; Acir and Guzelis, 2004; Pal and Mather, 2004; Huang *et al.*, 2004; Distant *et al.*, 2003). SVM are a learning algorithm that is based on the statistical learning theory. It was developed by Vapnik (1999) and uses the structural risk minimisation principle. Given an input vector, the SVM classifies it into one of two classes. This is done by seeking the optimal separating hyperplane that provides efficient separation for the data and maximises the margin, i.e. it maximises the distance between the closest vectors in both classes to the hyper plane. This is of course with the assumption that the data are linearly separable. For the case when the data are not linearly separable, SVM will map the data into a higher dimensional feature space where data can become linearly separable. This is done by using kernel functions. More details on SVM can be found in Qu *et al.* (2003) and Vapnik (1999)

3.3. RADIAL BASIS FUNCTIONS NETWORKS (RBFN)

RBFN are powerful interpolation techniques that can be efficiently applied in multidimensional space. The RBFN approach to classification is based on curve fitting. Learning is achieved when a multi-dimensional surface is found that can provide optimum separation of multi-dimensional training data. In general, RBFN can model continuous functions with reasonable accuracy. The radial basis functions are the set of functions provided by the hidden nodes that constitute an arbitrary “basis” for the input patterns (Qu *et al.*, 2003).

The training of the RBFN is usually simpler and shorter compared to other NNs. This is one of major advantages for using RBFN. However, greater computation and storage requirements for classification of inputs are usually required after the network is trained (Sutton and Barto, 1998).

RBFN apply a mixture of supervised and unsupervised learning modes. The layer from input nodes to hidden nodes is unsupervised, while supervised learning exists in the layer from hidden nodes to output nodes. Non linear transformation exists from input to hidden space, while linear transformation exists from the hidden to the output space (Bishop, 1995). More details on the theory and implementation of RBFN can be found in Qu *et al.* (2003) and Sutton and Barto (1998)

4. Implementation and Results

4.1 ASSOCIATING SUNSPOTS AND FLARES

As illustrated previously, a C++ platform was created to access the NGDC sunspots and flares catalogues and to automatically associate all the X and M class flares with their corresponding sunspot groups. The correspondence was determined based on the location (i.e., same NOAA number) and timing information (maximum six hours time difference between the erupting flare and the time for which the classification of the associated sunspot group is reported). After finding all the associations, a numerical dataset is created for the machine learning algorithms.

The whole process is described in the following steps:

- Parse flare data from NGDC soft X-ray flare catalogues.
- Parse sunspot group data from NGDC sunspot catalogues.
- If an X- or M-class flare is found then try to identify its associated sunspot as follows:
 - If the NOAA number of the flare is specified in the flare catalogue then,
 - Search the sunspot catalogue to find the sunspot groups with the matching NOAA number with flare.
 - Find the time difference between the flare eruption time and the extracted sunspots time.
 - If the time difference is less than six hours then the flare record and the sunspot record are marked as being ASSOCIATED.
 - If there is more than one classification report for the flare-associated sunspot group within six hours then choose the one with minimum time difference.
- Get the sunspot groups that are not associated with any flare.
 - If no solar flares anywhere on the sun occur within one day after the classification of this sunspot group then it is marked as UNASSOCIATED.
- Record all ASSOCIATED groups and their corresponding flare classes (An example for ASSOCIATED groups and flares is given in Figure 1).
- Record UNASSOCIATED groups.
- Create the data sets using McIntosh classes of groups and flare classes (Table I).

4.2 CREATING THE INITIAL DATA SET

We need to define appropriate numerical representations for the ASSOCIATED and UNASSOCIATED groups that were extracted earlier, because most machine learning algorithms deal with numbers. This means that we need to convert the McIntosh classification of every sunspot to a corresponding numerical value. To do so we will go back to the basics of McIntosh classification which is represented as Zpc , as explained earlier. There are seven classes under the Z part of classification (i.e., A, B, C, D, E, F, H), six classes under the p part (i.e., x, r, s, a, h, k) and four classes under the c part (i.e.,

x, o, i, c). Hence, three numerical values are used to represent every sunspot group. To find these values the McIntosh classification for the sunspot group is divided into its three components and the class order is divided by the total number of classes for each component. As for the flare representation, two numerical values are used. The first value represents whether a significant flare occurs or not. This value becomes 0.9 if a significant flare occurs otherwise it is 0.1. Also if the flare is an X-class flare, it is represented by 0.9 and if it is an M-class flare, it is represented by 0.1. The numerical conversions of data sets are explained in more details in Table I.

In this work, we have processed all the data of sunspots and flares for the periods from 01 Jan1992 till 31 Dec 2005. Our software has analysed the data related to 29343 flares and 110241 sunspot groups and has managed to associate 1425 X- or M-class flares with their corresponding sunspot groups. To complete our data set we have processed the NGDC catalogues again to choose 1457 UNASSOCIATED sunspot groups based on the rules mentioned previously. Hence, the total number of samples used for our training and testing sets for machine learning is 2882.

It is worth mentioning here that the total number of reported X- or M-class flares in NGDC flares catalogues is 1870 which is 6.3% of the total numbers of flares in this catalogue. However, out of these 1870 flares, only 1425 flares have NOAA numbers and are included in our study. Veronig *et al.* (2002) has reported that about 10% of all soft X-ray flares are X- or M-class flares based on studying all flares between 1976 and 2000. Comparing these findings with ours reveals that there is a 3.7% difference. The data used for Veronig *et al.* (2002) is more comprehensive as it consisted of 24 years data with three solar maxima, while we have extracted all the available NGDC data which is fourteen years data with only one solar maximum. Veronig *et al.* (2002) commented that “M- and X-class flares are less frequent and occur primarily during times of maximum solar activity” which may explain the 3.7% difference between the two findings.

4.3 THE INITIAL TRAINING AND TESTING PROCESS

4.3.1. *Jack-Knife Technique*

For our study, 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. Practically, a random number generator decides 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 classifier (Fukunaga, 1990). In each experiment, 80% of the samples were randomly selected and used for training while the remaining 20% were used for testing. Hence, the number of training samples is 2305, while 577 samples are used for the testing of the classifier.

4.3.2. *Initial Training and Testing*

To test the suitability of our data set for solar flare prediction, we started our experiments with training and testing of CCNN with our initial data set. This data set is divided into inputs and outputs as described in Section 4.2. The input set of our training

data consisted of three numerical values representing the McIntosh classification of the sunspot. The output set consisted of two nodes representing whether a flare is likely to occur and whether it is an X- or M-class flare. Several experiments based on the Jack-knife technique were carried out and we found that the prediction rate for flares in the best case scenario was 72.9%. This indicated that a correlation existed between the input and output sets, but it is not high enough to provide reliable prediction of solar activities. It soon became obvious that our training data do not enable the classifier to create accurate discriminations between the input and output data. We concluded that using the McIntosh classification on its own is not enough and more discriminating data need to be added to the input data. Hence, we decided to associate the classified sunspots with the sunspot cycle. The new input is computed by finding the time of the sunspot with respect to the solar cycle. Our decision to include the timing information seemed to be supported by the fact that the rise and fall of solar activity coincides with the sunspot cycle (Pap *et al.*, 1990) and by the fact that the M- and X- class flares mostly occur during times of maximum solar activity (Veronig *et al.*, 2002). When the solar cycle is at a maximum, plenty of large active regions exist and many solar flares are detected. The number of active regions decreases as the Sun approaches the minimum part of its cycle (Pap *et al.*, 1990). More details are provided in the section to follow.

4.3.3. Sunspot Cycle

To determine the association between the classified sunspot group and their location on the solar cycle, it is important to use a mathematical model that could simulate the behaviour of the solar cycle. The model introduced by Hathaway *et al.* (1994), which is shown in Equation (1), can serve this purpose. This equation enables us to estimate the sunspot number for any given date and it enables the prediction of future values as well.

$$f(t) = a(t - t_0)^3 / \{ \exp[(t - t_0)^2 / b^2] - c \} \quad (1)$$

In Equation (1), parameter ‘ a ’ represents the amplitude and is related to the rise of the cycle minimum, ‘ b ’ is related to the time in months from minimum to maximum; ‘ c ’ gives the asymmetry of the cycle; and ‘ t_0 ’ denotes the starting time. More information about the derivation of Equation (1) can be found in Hathaway *et al.* (1994).

To verify the accuracy of this model we have applied this model to the dates from 01 Jan 1992 till 31 Dec 2005 and the average of the sunspot numbers for each month is shown in Figure 2. Using our software, we have extracted the sunspot data for the same period and found the actual average of monthly sunspot numbers. This value is also shown in Figure 2. Comparing both plots, it is obvious that Hathaway’s model provides reliable modelling for the rise and fall of the solar cycle. An alternative method to simulate the solar cycle could be the NN model introduced by Calvo *et al.* (1995). It is worth mentioning that in Lantos (2006) it is shown that all solar cycle prediction methods give, at least for one of the cycles, an error larger than 20 %, especially when cycles 13 to 23 are considered. Although Hathaway’s model represents the evolution of the solar cycle in a slightly rough manner, we used this model because of its easy implementation for the existing and future solar cycles. In addition, Hathaway’s model provides an automated numerical simulation for the behaviour of the solar cycle that can be used for studying the past, present and future evolution of the cycle. This is particularly important for any automated prediction system as certain sunspots group classes are more likely to flare at certain times. Hence, Hathaway’s model is integrated

into the training and testing phases of our machine learning algorithms. After the machine learning algorithms are trained, the timing information for any future sunspot can be simulated and fed to them to provide real-time flares prediction.

4.3.4. *The Final Data Set*

After using Hathaway's model, the final training data set can be constructed. For each sample, the training data set consists of six numerical values: four input values and two output values. The first three input values represent the McIntosh classification of sunspots, while the last numerical value represents the simulated and normalized sunspot number based on Hathaway's model. The simulated Hathaway's value is normalised by dividing it by the actual or estimated number of sunspots at solar maxima to obtain a value between zero and one. The output set consists of two values: The first value is used to predict whether the sunspot is going to produce a significant flare or not. For performance evaluation purposes, this value will be referred to as the "Correct Flare Prediction (CFP)" value. The second value is used to determine whether the predicted flare is an X or M flare, and it will be referred to as the "Correct Flare Type Prediction (CFTP)". The final input and output data set is given in Table II. Also some examples from the data set are provided in Table III.

4.4. FINAL TRAININGS AND TESTINGS FOR COMPARISONS

In order to determine which machine learning algorithm is more suitable for our application, the prediction performance of CCNN, SVM and RBFN needs to be compared. 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. 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 other learning algorithms.

4.4.1 *Optimising the CCNN for final comparison*

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.

4.4.2 *Optimising the SVM for final comparison*

For our work, the SVM experiments are carried out using the mySVM program developed by S.Ruping². To optimise the performance of SVM the optimal kernel and its parameters need to be determined empirically. There are no known guidelines to choose them (Manik *et al.*, 2004). In our work, many kernels were tested, such as the

² <http://www-ai.informatik.uni-ortmund.de/SOFTWARE/MYSVM/index.html>.

dot, polynomial, neural, radial, and Anova kernels, in order to choose the kernel that provides the best classification performance. After applying all these kernels we have concluded that the Anova kernel, which is shown in Equation (2), provides better results compared to all other kernels. Our findings match the findings of Manik *et al.* (2004).

$$k(x, y) = \left\{ \sum_i \exp[-\gamma(x_i - y_i)] \right\}^d \quad (2)$$

In Equation (2), the sum of exponential functions in x and y directions defines the Anova kernel. The shape of the kernel is controlled by the parameters γ (gamma) and d (exponential degree). To complete our SVM optimisation we need to determine these parameters for the Anova kernel. Two sets of experiments are carried out to determine these values. In the first experiments the value of γ is fixed at 50 and several SVM training experiments are concluded while changing the value of d from one to ten. For every new experiment the value of d is incremented and the prediction performance is recorded. The results for CFP and CFTP for different d values are shown in Figure 5. This figure shows that saturation is reached as the value of d exceeds eight. In the second experiments d is fixed at five and the value of γ is changed from 10 to 100. Figure 6 shows the results for CFP and CFTP for different γ values. The best results are reached for $\gamma = 85$. Based on the above, the Anova kernel is used with its d and γ parameters set to 8 and 85 consecutively.

4.4.3 Optimising the RBFN for final comparison

A standard RBFN has a feed-forward structure consisting of three layers, an input layer, a nonlinear hidden layer and a linear output layer (Broomhead and Lowe, 1988). We have used MATLAB's RBFN tool for our experiments. This tool adds nodes to the hidden layer of a RBFN until it reaches the specified learning goal, which is usually the mean squared error value. For our case this value is made equal to 0.001. To optimise the RBFN, we have run many simulations by increasing the number of hidden nodes from 3 to 60 in order to determine the optimum RBFN topology. Figure 7 shows the comparison for RBFN simulations with different number of hidden nodes. It is clear that the optimum number of hidden nodes is between 20 and 40 nodes. Hence, we have decided to use 30 hidden nodes for our future experiments.

4.5. COMPARING THE PERFORMANCES

After successfully optimising CCNN, SVM and RBFN, we are able to compare the prediction performance for the three learning algorithms. As illustrated previously, our CCNN has two hidden layers with six and four hidden nodes, our SVM uses the Anova kernel with d and γ parameters set to 8 and 85 respectively and our RBFN has 30 hidden nodes. All the training and testing experiments are carried out based on the statistical Jack-knife technique. For every experiment the Jack-knife technique is applied once to obtain the training and testing sets. For every experiment the same data sets are shared between CCNN, SVM and RBFN. Ten experiments are carried out and the results for CFP, CFTP, and training times are shown in Table IV.

As it can be seen from Table IV, CCNN requires more training times compared to SVM and RBFN. On the other hand, SVM is very fast in training and on average it provides the best performance for CFP, which is 93.07%. The best performance for CFTP is provided by CCNN, which on average is 88.02%. It is worth noting that RBFN

outperforms SVM for CFTP as well. Figure 8 and Figure 9 shows the graphical comparison of CFP and CFTP performances for the three learning algorithms.

5. Conclusions and Future Research

In this paper, an empirical evaluation for three of the top machine learning algorithms for the automated prediction of solar flares is carried out. CCNN, RBFN and SVM are compared based on their topology, training times and most importantly their prediction accuracy. We used the publicly available solar catalogues from the National Geophysical Data Centre to associate the reported occurrences of M and X solar flares with the relevant sunspots that were classified manually and exist in the same NOAA region prior to flares occurrence.

We have investigated all the reported flares and sunspots between 01 Jan 1992 and 31 Dec 2005. Our software has managed to associate 1425 M and X soft X-ray flares with their corresponding sunspot groups. These 1425 flares were produced by 446 different sunspot groups. Among these groups, NOAA region 10808 produced 30, region 10375 produced 29 and region 9393 produced 27 of the major flares. Our software was used to extract these findings and it can be used for statistical-based analysis of different aspects connected to flares occurrence.

Our practical findings on the associations show that there is a direct relation between the eruptions of flares and certain McIntosh classes of sunspots such as Ekc, Fki and Fkc. Our findings are in accordance with McIntosh (1990), Warwick (1966), and Sakurai (1970). However we believe that our work is the first to introduce a fully automated computer platform that could verify this relation using machine learning. In its current version, our system can provide a few hours prediction in advance (i.e., up to six hours in advance) by analysing the latest sunspots data.

We compared different machine learning algorithms to obtain the best prediction results for solar flares. Erupting flares were associated with their corresponding sunspot groups with McIntosh classifications. To increase the accuracy of prediction, a mathematical model representing the sunspot cycle based on the work of Hathaway *et al.* (1994), is implemented to automatically simulate the solar cycle. The simulated solar cycle and classified sunspots are converted to the appropriate numerical formats and used to train CCNN, RBFN and SVM machine learning systems, to predict automatically whether a significant flare will occur and whether it is going to be an X- or M-class flare.

In this research, SVM provides the best results for predicting if a sunspot group with McIntosh classification is going to flare or not but CCNN is more capable of predicting the class of the flare to erupt. We have reached the conclusion that a hybrid system similar to the one shown in Figure 10, which combines both SVM and CCNN, will give better results for flare prediction. In this hybrid system, the inputs will be fed to SVM engine to decide if the sunspot group is going to produce a solar flare or not. Depending on the outcome of the SVM engine, CCNN is used to predict the type of the flare to erupt.

We believe that this work is a first step toward constructing a fully automated and web-compliant platform that would provide short-term prediction for solar flares. In this paper, we managed to extract the experts' knowledge which is embedded in the catalogues of flares and sunspots and managed to represent this knowledge using association and learning algorithms. However, the creation of an automated and objective system means that the system should be able to accept the latest solar images, detect solar features (i.e., sunspots, active regions, etc), provide the McIntosh classification for the detected features and then feed these classifications to our hybrid

system to obtain the short-term prediction for solar activities. This would require the use of advanced image processing techniques and we are currently working on this.

Another future direction is to incorporate the findings of Ternullo *et al.* (2006) and modify our association software to extract the age of every associated sunspot. A flaring probability that increases with the age of this sunspot can be added to our input features. This could help in increasing the prediction performance even further. The association software can also be modified to study the development of flaring sunspots. Sunspots are dynamic objects and they evolve with time. Most of the existing prediction models don't take this evolution into consideration when predicting flares (Wheatland, 2004). A third direction is to combine the Mt Wilson classification of sunspots with McIntosh classification. This is inspired by the findings of Sammis *et al.* (2000), where it was proven that there is a general trend for large sunspots to produce large flares, but the most important factor is the magnetic properties of these sunspots. A serious problem facing these future directions is that the observations of sunspots are not made at fixed time intervals. We need to find a solution to this problem to improve the consistency of these data and make it more suitable for machine learning algorithms.

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

- Acir, N. and Guzelis, C.: 2004, *Expert Systems with Applications* **27**, 451.
- Benkhalil, A., Zharkova, V., Ipson, S., and Zharkov, S.: 2003, in H Holstein and F. Labrosse (eds.), *AISB'03 Symposium on Biologically-Inspired Machine Vision, Theory and Application*, University of Wales, Aberystwyth, p. 66.
- Bishop, C.M.: 1995, *Neural Networks for Pattern Recognition*, Oxford University Press, London, p.164.
- Borda, R.A.F., Mininni, P.D., Mandrini, C.H., Gomez, D.O., Bauer, O.H., and Rovira, M.G.: 2002, *Solar Phys.* **206**, 347.
- Bornmann, P.L. and Shaw, D.: 1994, *Solar Phys.* **150**, 127.
- Broomhead, D.S. and Lowe, D.: 1988, *Complex Systems* **2**, 321.
- Calvo, R.A., Ceccatto, H.A., and Piacentini, R.D.: 1995, *Astrophys. J.* **444**, 916.
- Distante, C., Ancona, N., and Siciliano, P.: 2003, *Sensors and Actuators B: Chemical* **88**, 30.
- Fahlmann, S.E. and Lebiere, C.: 1989. in D. S. Touretzky(ed.), *Advances in Neural Information Processing Systems 2 (NIPS-2)*, Morgan Kaufmann, Denver, Colorado, p. 524.
- Frank, R.J., Davey, N., and Hunt, S.P.: 1997. *J. Intelligent and Robotic Systems*, **31**, 91.
- Fukunaga, K.: 1990, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, p. 220.

- Gallagher, P.T., Moon, Y.J., and Wang, H.M.: 2002, *Solar Phys.* **209**, 171.
- Gao, J.L., Wang, H.M., and Zhou, M.C.: 2002, *Solar Phys.* **205**, 93.
- Greatrix, G.R. and Curtis, G.H.: 1973, *Observatory* **93**, 114.
- Hale, G.E., Ellerman, F., Nicholson, S.B., and Joy, A.H. 1919, *Astrophys. J.* **49**, 153.
- Hathaway, D., Wilson, R.M., and Reichmann, E.J.: 1994, *Solar Phys.* **151**, 177.
- Huang, Z., Chen, H.C., Hsu, C.J., Chen, W.H., and Wu, S.S.: 2004, *Decision Support Systems* **37**, 543.
- Koskinen, H., Eliasson, L., Holback, B., Andersson, L., Eriksson, A., Mälkki, A., Norberg, O., Pulkkinen, T., Viljanen, A., Wahlund, J.-E., and Wu, J.-G.: 1999. *Space Weather and Interactions with Spacecraft, Finnish Meteorological Institute Reports 1999-4*.
- Künzel, H.: 1960, *Astron. Nachr.* **285**, 271.
- Lantos, P.: 2006, *Solar Phys.* **236**, 199.
- Lefebvre, S. and Rozelot, J.P.: 2004, *Solar Phys.* **219**, 25.
- Liu, C., Deng, N., Liu, Y., Falconer, D., Goode, P.R., Denker, C., and Wang, H.M.: 2005, *Astrophys. J.* **622**, 722.
- Manik, A., Singh, A., and Yan, S.: 2004, in E.G. Berkowitz (ed.), *Fifteenth Midwest Artificial Intelligence and Cognitive Sciences Conference*, Omnipress, Chicago, p.74.
- McIntosh, P.S.: 1990, *Solar Phys.* **125**, 251.
- Pal, M. and Mather, P.M.: 2004, *Future Generation Computer Systems* **20**, 1215.
- Pap, J., Bouwer, S., and Tobiska, W.: 1990, *Solar Phys.* **129**, 165.
- Qahwaji, R. and Colak, T.: 2006a, *Int. J. Computers and Their Appl.* **13**, 9.
- Qahwaji, R. & Colak, T.: 2006b, in H.W. Chu, J. Aguilar, N. Rishe, and J. Azoulay (eds.), *The Third International Conference on Cybernetics and Information Technologies, Systems and Applications*, International Institute of Informatics and Systemics, Orlando, p.192.
- Qu, M., Shih, F.Y., Jing, J., and Wang, H.M.: 2003, *Solar Phys.* **217**, 157.
- Sakurai, K.: 1970, *Planet Space Sci.* **18**, 33.
- Sammis, I., Tang, F., and Zirin, H.: 2000, *Astrophys. J.* **540**, 583.
- Schetinin, V.: 2003, *Neural Processing Lett.* **17**, 21.
- Shet, R.N., Lai, K.H., A.Edirisinghe, E., and Chung, P.W.H.: 2005, in J.S. Marques, N. Pe'rez de la Blanca, and P.Pina (eds.), *Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, Vol. 3523, Springer Verlag, Berlin, p.343.
- Shi, Z.X. and Wang, J.X.: 1994, *Solar Phys.* **149**, 105.

- Shih, F.Y. and Kowalski, A.J.: 2003, *Solar Phys.* **218**, 99.
- Smieja, F.J.: 1993, *Circuits, Systems and Signal Processing* **12**, 331.
- Sutton, R.S. and Barto, A.G.: 1998, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Massachusetts, p.193.
- Ternullo, M., Contarino, L., Romano, P., and Zuccarello, F.: 2006, *Astron. Nachr.* **327**, 36.
- Turmon, M., Pap, J.M., and Mukhtar, S.: 2002, *Astrophys. J.* **568**, 396.
- Vapnik, V.: 1999, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, p. 138.
- Veronig, A., Temmer, M., Hanslmeier, A., Otruba, W., and Messerotti, M.: 2002, *Astron. Astrophys.* **382**, 1070.
- Wang, H., Qu, M., Shih, F., Denker, C., Gerbessiotis, A., Lofdahl, M., Rees, D. and Keller, C.: 2003, *Bull. Am. Astron. Soc*, **36**, 755.
- Warwick, C.S.: 1966, *Astrophys. J.* **145**, 215.
- Wheatland, M.S.: 2004, *Astrophys. J.* **609**, 1134.
- Zharkova, V., Ipson, S., Benkhalil, A., and Zharkov, S.: 2005, *Artificial Intelligence Review*, **23**, 209.
- Zharkova, V. and Schetin, V.: 2003, in L. C. Jain, R. J. Howlett, and V. Palade (eds.), *Proceedings of the Seventh International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES'03*, Springer, Oxford, U.K., p.148.
- Zirin, H. and Liggett, M.A.: 1987, *Solar Phys.* **113**, 267.

Figure 1: Parts from NGDC flare and sunspot catalogues for 02/Nov/2003 showing an associated flare and sunspot region.

Figure 2: Comparison of the actual and simulated ‘monthly average sunspot number’ between 01 Jan 1992 and 31 Dec 2005.

Figure 3: Comparison of CCNNs with different number of hidden nodes for CFP.

Figure 4: Comparison of CCNNs with different number of nodes for CFTP.

Figure 5: Comparison of SVM with different degree values.

Figure 6: Comparison of SVM with different gamma values.

Figure 7: Comparison of RBFN with different number of maximum hidden nodes.

Figure 8: Graphical comparison of experiment results for Correct Flare Prediction (CFP) using different learning algorithms.

Figure 9: Graphical comparison of experiment results for Correct Flare Type Prediction (CFTP) using different learning algorithms.

Figure 10: Suggested hybrid flare prediction system.

Table I: Inputs and output values for initial machine training process.

Inputs			Outputs	
McIntosh classes			Flare	No/ M-class
A= 0.10	X= 0		=0.9	flare =
H= 0.15	R=0.10	X=0		0.1
B= 0.30	S=0.30	O=0.10	No	X-class
C= 0.45	A=0.50	I=0.50	flare=	Flare =
D= 0.60	H=0.70	C=0.90	0.1	0.9
E= 0.75	K=0.90			
F= 0.90				

Table II: Inputs and output values for final machine training process.

Inputs				Outputs	
McIntosh classes			Normalized average sunspot number from Hathaway function	Flare =0.9 No flare= 0.1	No/M-class flare = 0.1 X-class Flare = 0.9
A= 0.10 H= 0.15 B= 0.30 C= 0.45 D= 0.60 E= 0.75 F= 0.90	X= 0 R=0.10 S=0.30 A=0.50 H=0.70 K=0.90	X=0 O=0.10 I=0.50 C=0.90			

Table III: Some example input and output values used for machine training process: Data 1 is representing a X-class flare detected on 02/Nov/2003 at 17:39 on NOAA region 10486 which is classified as EKC at 15:45. Data 2 is representing a M-class flare detected on 02/May/2003 at 02:47 on NOAA region 10345 which is classified as DAO at 00:45. Data 3 is representing NOAA region 7566 which is classified as CSO at 08:40 and no corresponding flares detected at that day.

Data no	Inputs				Outputs	
1	E	K	C	Normalized sunspot number	Flare	X-class
	0.75	0.90	0.90	0.145165	0.9	0.9
2	D	A	O	Normalized sunspot number	Flare	M-class
	0.60	0.50	0.10	0.189032	0.9	0.1
3	C	S	O	Normalized sunspot number	No flare	No flare
	0.45	0.30	0.10	0.100992	0.1	0.1

Table IV: Experiments and results using Jack-knife technique. (CFP= Correct Flare Prediction, CFTP= Correct Flare Type Prediction, TT=Training Time in seconds).

Experiments	CCNN			SVM			RBFN		
	TT.	CFP	CFTP	TT.	CFP	CFTP	TT.	CFP	CFTP
1	32.14	92.72	88.91	0.05	93.41	82.50	13.94	89.77	85.96
2	31.83	93.07	88.39	0.03	94.28	83.54	13.41	89.95	85.44
3	31.72	90.47	86.31	0.05	91.33	82.15	13.75	88.21	84.06
4	31.94	93.24	89.43	0.03	93.07	81.98	13.92	90.64	86.83
5	31.73	92.20	88.73	0.05	94.11	86.48	13.70	88.39	84.92
6	31.75	90.12	85.96	0.02	93.07	84.40	13.34	89.25	85.10
7	31.63	92.20	89.25	0.03	93.93	84.92	13.34	90.47	87.52
8	32.00	91.85	88.91	0.06	93.93	84.92	13.38	90.47	87.52
9	32.00	91.16	88.39	0.03	92.20	84.23	13.66	88.21	85.62
10	31.78	88.91	85.96	0.05	91.33	82.15	13.66	89.08	85.96
Average	31.85	91.59	88.02	0.04	93.07	83.73	13.61	89.45	85.89