Automatic Prediction of Solar Flares using Machine Learning: Practical Study on the Halloween Storm

R.Qahwaji[1] , T. Colak[2]
[1]Department of Electronic Imaging and Media Communications, r.s.r.qahwaji@bradford.ac.uk
Horton A2.4 B, Richmond Road, Bradford BD7 1DP, England, U.K.
[2] Department of Electronic Imaging and Media Communications, t.colak@bradford.ac.uk
Chesham B1.14, Richmond Road, Bradford BD7 1DP, England, U.K.

Abstract- **In this paper, a machine learning system that can provide short-term automated prediction for the occurrence of significant solar flares is presented. This system extracts the experts' knowledge embedded in the public NGDC solar catalogues and represents it in learning rules that can be used by computers to predict flares. This work builds on our previous work and the prediction system is tested intensively using the Jackknife technique and using real input samples from the Halloween storm. The system has managed to predict all the significant flares that took place during this storm.**

I. Introduction

Severe solar activities can have significant impact on our life on Earth. The most dramatic solar activity events affecting the terrestrial environment are solar flares and Coronal Mass Ejections (CMEs). Flares and CMEs are two types of solar eruption that can spew vast quantities of radiation and charged particles into space. The Earth environment and geomagnetic activity is affected by the ionized solar plasma, also known as the solar wind. The solar wind flows outward from the Sun to form the heliosphere and it is affected by solar activity and carries with it the magnetic field of the Sun. This interplanetary magnetic field (IMF) creates storms by injecting plasma into the Earth's magnetosphere.

Solar activities can affect wireless communications systems causing interruption of service (e.g., frequency jamming and dropped communications) due to radio bursts caused by microwave emissions from the Sun. The arrival of solar X-rays that are traveling at the speed of light can disrupt point-to-point high frequency radio communications. This phenomenon, called a sudden ionospheric disturbance (SID), has been associated for a long time with solar flares. Concurrent with X-ray emission, solar activity often emits radio noise that can interfere with communications and radar systems on the sunlit side of Earth. Adverse space weather conditions can cause anomalies and system failures in spacecraft operations and in satellites. During periods of increased solar activity the outer atmosphere of the Earth expands, which results in a greater drag on the movement of satellites and spacecraft. This causes a slow-down and a change in orbit that could shorten the life-time of these missions. Energetic particles from the Sun may cause direct physical damage to the equipment and the spacecraft (e.g., solar panels). Low and high Earth-orbiting spacecraft and satellites are subject to a number of environmental radiation hazards, such as direct collision and/or electrical upsets, caused by charged particles. In addition, high-energy charged particles which result from major bursts of solar activity (i.e., CMEs) are hazardous to astronauts on space missions. In space, astronauts perform extra-vehicular activities (EVAs) and they can be subjected to solar energetic particle events and cosmic ray particles. Particle energies can increase hundreds of times after an extra ordinary solar flare and/or CME and can endanger the life of astronauts. The Halloween storm, which occurred late October early November 2003, caused serious problems including damaging 28 satellites, knocking two out of commission, diverting airplane routes and causing power failures in Sweden, and others.

Satellite operators, space agencies, aviation industry, power generation and distribution industry, oil and gas industry and railways can benefit from an effective space weather prediction system. The importance of space weather will continue to increase because our reliance on satellites for communications and resource information (meteorological, geophysical prospecting, navigation, and remote sensing) increases year after year. The ability to predict major solar storms can give organizations sufficient lead time to implement preventive and safety measures.

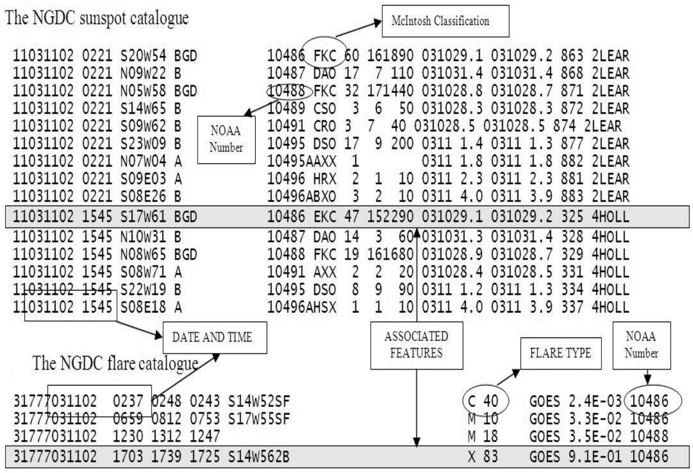
The aim of this paper is to provide a practical example on how machine learning can be used to predict the actual occurrence of large solar flares. The fully automated prediction system could extract the experts' knowledge that is embedded in solar data catalogues to provide short-term real-time prediction for solar flares. This knowledge is extracted and represented in terms of learning rules that could be processed by computers using machine learning techniques. Many experiments are carried out using the Jackknife technique to study the efficiency of this system. The system is also tested on the sunspot groups that are available before and during the Halloween storm of 2003. The predictions of our system will be compared with the actual flare eruptions that took place then. This paper is organized as follows: Section II provides information about the public solar catalogues used in this paper. The machine learning algorithm is described in Section III. Section IV is devoted to the association of solar features, practical implementation and the testing of our algorithm. Finally, the concluding remarks are provided in Section V.

II. DATA

Data from the publicly available sunspot group catalogue and the solar flare catalogue, which is provided by the National Geophysical Data Centre (NGDC), is used for this work. 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 1 to 8 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. This catalogue provides information about dates, starting and ending times for flare eruptions, location, NOAA number of the corresponding active region and x-ray classification for the detected flares. Not all the flares have associated NOAA numbers. Flares without NOAA numbers are not included in our study. A sample of the NGDC catalogues used in our work is shown in Fig. 1.

The NGDC sunspot catalogue holds records of sunspot groups supplying their location, time, physical properties and classification data. 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 [1] is based on the distribution of magnetic polarities within spot groups [2]. The McIntosh classification is the standard for the international exchange of solar geophysical data. It is a modified version of the Zurich classification system developed by Waldmeir. The general form of the McIntosh classification is *Zpc* where, *Z* is the modified Zurich class, *p* is the type of spot, and *c* is the degree of compactness in the interior of the group. Mt. Wilson classification consists of letters taken from the Greek alphabet from alpha to delta and their different combination.


The NGDC sunspot catalogue
McIntosh Classification
NOAA Number
DATE AND TIME
ASSOCIATED FEATURES
FLARE TYPE
NOAA Number
The NGDC flare catalogue

```
11031102 0221 S20W54 BGD        10486 FKC 60 161890 031029.1 031029.2 863 2LEAR
11031102 0221 N09W22 B          10487 DAO 17   7 110 031031.4 031031.4 868 2LEAR
11031102 0221 N05W58 BGD        10488 FKC 32 171440 031028.8 031028.7 871 2LEAR
11031102 0221 S14W65 B          10489 CSO  3   6  50 031028.3 031028.3 872 2LEAR
11031102 0221 S09W62 B          10491 CRO 3  7   40 031028.5 031028.5 874 2LEAR
11031102 0221 S23W09 B          10495 DSO 17   9 200 0311 1.4 0311 1.3 877 2LEAR
11031102 0221 N07W04 A          10495AAXX  1         0311 1.8 0311 1.8 882 2LEAR
11031102 0221 S09E03 A          10496 HRX  2   1  10 0311 2.3 0311 2.3 881 2LEAR
11031102 0221 S08E26 B          10496ABXO  3   2  10 0311 4.0 0311 3.9 883 2LEAR
11031102 1545 S17W61 BGD        10486 EKC 47 152290 031029.1 031029.2 325 4HOLL
11031102 1545 N10W31 B          10487 DAO 14   3  60 031031.3 031031.4 328 4HOLL
11031102 1545 N08W65 BGD        10488 FKC 19 161680 031028.9 031028.7 329 4HOLL
11031102 1545 S08W71 A          10491 AXX  2   2  20 031028.4 031028.5 331 4HOLL
11031102 1545 S22W19 B          10495 DSO  8   9  90 0311 1.2 0311 1.3 334 4HOLL
11031102 1545 S08E18 A          10496AHSX  1   1  10 0311 4.0 0311 3.9 337 4HOLL

31777031102  0237 0248 0243 S14W52SF                  C 40   GOES 2.4E-03 10486
31777031102  0659 0812 0753 S17W55SF                  M 10   GOES 3.3E-02 10486
31777031102  1230 1312 1247                           M 18   GOES 3.5E-02 10488
31777031102  1703 1739 1725 S14W562B                  X 83   GOES 9.1E-01 10486
```

Fig. 1. A sample of the NGDC sunspots and flares catalogues

III. THE MACHINE LEARNING ALGORITHM

In this work, Cascade-Correlation Neural Networks (CCNNs) are used for flare prediction. CCNNs are used because of their efficient performance in applications involving classification and time-series prediction [3]. In our previous work [4] 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. A full comparison between the prediction performances for CCNN, Support Vector Machines (SVM) and Radial Basis Function Networks (RBFN) is carried out in our recent work [5].

A. Cascade-Correlation Neural Networks (CCNNs)

The training of Backpropagation neural networks is considered to be slow because of the step-size problem and the moving target problem [6]. 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, hence making it a permanent feature-detector in the network. This feature detector can then produce outputs or create other more complex feature detectors [6].

In a 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 [6].

However, a major problem facing these networks is over-fitting the training data, especially when dealing with real-world problems [7]. Over-fitting usually occurs if the training data are characterized by many irrelevant and noisy features [8].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 [9]. Secondly, the network determines its own size and topology and it retains the structures it has built even if the training set changes [6]. Thirdly, it requires no back-propagation of error signals through the connections of the network [6]. Finally, this structure is useful for incremental learning in which new information is added to the already trained network [9]

IV. IMPLEMENTATION AND RESULTS

A. The Association and the Pre-Training Processes

For the training process, we have investigated all the sunspot groups that were associated with flares from 01 Jan1992 till 31 Dec 2006. The degree of association was determined based on the NOAA region number and the timing information. We have created a computer platform using C++ that would automatically access and extract information from the NGDC sunspot and flare catalogues. Our software has analysed the data related to 30683 flares and 111648 sunspots and has managed to associate 1450 M and X flares with their corresponding sunspot groups.

Theoretically, the total number of samples used for our training should be in the range of 2900 samples, where 1450 samples represent flaring features and the remaining 1450 samples representing non-flaring features. The flaring data represents the classifications and timing information of sunspots that produced flares. The remaining samples represent sunspots that existed in non-flaring days and are not related to any sunspot groups within the previous flaring sunspot samples together with their timing information. However, we have decided to exclude the samples corresponding to the period of 01 July 2003 till 31 Dec 2003 to provide an additional test to whether our prediction system can successfully predict the occurrence of the Halloween flares. The number of excluded flaring samples is 76. Hence, a similar number of non sampling features is also excluded. This means that the total number of training samples has been reduced to 2748.

All the machine learning training and testing experiments are carried out with the aid of the Jack-knife technique [10]. 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 2 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 [10]. In each experiment, 80% of the samples remaining after excluding the 76 samples described above, are randomly selected and used for training while the remaining 20% are used for testing. Hence, the number of training samples is 2198, while 550 samples are used for the testing of the classifier. Keeping in mind that some samples are deliberately excluded, this means that the learning system is trained only with 75.8% of the total number of associated samples.

B. The training Vectors

For each sample, the training vector consists of 6 numerical values that belong to two sets: the input set and the target set. The input set contains 4 values, while the remaining two values belong to the output set. The first three values of the input set represent the sunspot McIntosh classification, while the last numerical value represents the simulated sunspot number which is generated based on Hathaway's model [11] for the given dates. The three McIntosh classification values are the modified Zurich class, type of largest spot and the sunspot distribution. On the other hand, the target set consists of two values representing whether a flare is likely to occur and whether it is an X or M flare.

C. Optimising the CCNN

In general, learning algorithms are optimised to ensure that their best performances are achieved. In [4], it was proven that CCNN provides the optimum neural network performance for processing catalogues data and in [5] it was shown that a CCNN with 6 hidden nodes in the first layer and 4 hidden nodes in the second layer provides the best results for Correct Flare Prediction (CFP) and Correct Flare Type Prediction (CFTP). This topology is used in this work.

D. Practical Experiments:

Two sets of testing experiments are carried out in this work. As explained earlier, the training stage for all experiments is carried out with only 75.8% of the total number of associated samples. For the first experiment, all the training and testing experiments are carried out based on the statistical Jack-knife technique. For every experiment the Jackknife technique is applied once to obtain the random training and testing sets. Ten experiments are carried with random samples representing the input and output features. For all the experiments the number of random training samples used is 2198. The remaining 550 samples are used for the testing of the classifier in the first experiment, while the 76 samples corresponding to the period from 01 July 2003 till 31 Dec 2003 and containing the samples for the Halloween storm are used for the second experiment. The results for training times, CFP and CFTP for the first and second experiments are shown in Table I and Table II, respectively.

As it can be seen from Table I, CCNN requires short training times and it provides high CFP rate for the short-term prediction of significant flares. The CFTP rate is slightly lower and we will be exploring more machine learning techniques in the future to increase its reliability.

TABLE I
PRACTICAL RESULTS FOR THE FIRST EXPERIMENT

Experiments	Training Time (Sec)	CFP	CFTP
1	41.3	0.927	0.909
2	39.3	0.916	0.878
3	39.1	0.916	0.878
4	39.3	0.913	0.878
5	39.4	0.913	0.878
6	39.6	0.933	0.898
7	39.7	0.916	0.889
8	40.4	0.924	0.871
9	40.9	0.907	0.867
10	41.1	0.931	0.898
Average	40.0	0.920	0.885

TABLE II
PRACTICAL RESULTS FOR THE SECOND EXPERIMENT

Experiments	Training Time (Sec)	CFP	CFTP
1	45.4	0.931	0.896
2	40.4	0.915	0.871
3	41.4	0.925	0.884
4	41.8	0.909	0.875
5	40.7	0.905	0.876
6	40.6	0.922	0.884
7	40.7	0.909	0.860
8	40.7	0.922	0.882
9	40.9	0.918	0.887
10	41.2	0.925	0.896
Average	41.4	0.918	0.881

It is worth mentioning that most of the major flares that occurred during the Halloween storm were associated with NOAA region 10486 such as: the X17.2 on 28th Oct 03, the X10 on 29th Oct 2003, the X8.3 on 2nd Nov 2003 and the X28 on 4th Nov 2003. All these flares are successfully predicted by our system.

V. CONCLUSIONS

In this work, we have 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 processed 30683 flares and 111648 sunspots between 01 Jan 1992 and 31 Dec 2006 and managed to associate 1450 M and X flares with their corresponding sunspot groups. To provide further testing we have excluded 76 samples between 01 July 2003 and 31 Dec 2003 to provide an additional test to whether our prediction system can successfully predict the occurrence of the Halloween flares. The knowledge embedded in the associated samples is extracted using the machine learning system that we have introduced here and represented in terms of learning rules that can be easily interpreted and applied by computers. These learning rules are the corner stone for the automated real-time system that we have introduced here for the prediction of significant solar flares. Extensive experiments using the Jack-knife technique are applied to evaluate the prediction performance of this system. The system is also tested on sunspots data belonging to the Halloween storm of 2003. Accurate and reliable performance is obtained for both testing experiments, as shown in Tables I and II.

Currently we are working on enhancing our association system so that it could extract the development, life cycle and age of every associated sunspot and then feed this information to the classifiers to enhance the overall prediction performance.

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

[1] G. E. Hale, F. Ellerman, S. B. Nicholson, and A. H. Joy, "The Magnetic Polarity of Sun-Spots," *Astrophysical Journal*, vol. 49, pp. 153, 1919.

[2] G. R. Greatrix and G. H. Curtis, "Magnetic Classification Of Sunspot Groups," *Observatory*, vol. 93, pp. 114-116, 1973.

[3] R. J. Frank, N. Davey, and S. P. Hunt, "Time Series Prediction and Neural Networks." *Journal of Intelligent and Robotic Systems*, pp. 91-103, 1997.

[4] R. Qahwaji and T. Colak, "Neural Network-based Prediction of Solar Activities," in *CITSA2006*. Orlando, 2006.

[5] R. Qahwaji and T. Colak, "Automatic Short-Term Solar Flare Prediction Using Machine Learning and Sunspot Associations," *Solar Physics*, 2007.

[6] S. E. Fahlmann and C. Lebiere, "The cascade-correlation learning architecture," presented at. In Advances in Neural Information Processing Systems 2 (NIPS-2), Denver, Colorado, 1989.

[7] F. J. Smieja, "Neural network constructive algorithms: Trading generalization for learning efficiency," *Circuits, Systems and Signal Processing*, vol. 12, pp. 331-374., 1993.

[8] V. Schetinin, "A Learning Algorithm for Evolving Cascade Neural Networks," *Neural Processing Letter*, vol. 17, pp. 21-31, 2003.

[9] R. N. Shet, K. H. Lai, E. A.Edirisinghe, and P. W. H. Chung, "Use of Neural Networks in Automatic Caricature Generation: An Approach Based on Drawing Style Capture," presented at The IEE International conference VIE2005 in Visual Information Engineering: Convergence in Graphic and Vision, Glasgow, UK, 2005.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition,” Academic Press, New York, 1990.* New York: Academic Press, 1990.

[11] D. Hathaway, R. M. Wilson, and E. J. Reichmann, "The Shape of the Sunspot Cycle," *Solar Physics*, vol. 151, pp. 177, 1994.