

On Learning of Sunspot Classification

Trung Thanh Nguyen¹, Claire P. Willis¹, Derek J. Paddon¹, Hung Son Nguyen²

¹ Department of Computer Science, University of Bath
Bath BA2 7AY, United Kingdom

² Institute of Mathematics, Warsaw University,
Banacha 2, Warsaw 02-095, Poland

Abstract. This paper describes automatic sunspot recognition and classification from satellite images. Some experimental results on learning sunspot classification using data mining techniques are presented. The classification scheme attempted was the seven-class Modified Zurich scheme. Daily images of the solar disk were taken from the NASA SOHO satellite's MDI instrument and sunspots extracted using image processing techniques. Two data mining tools, WEKA and RSES, were used for learning sunspot classification. In the training dataset sunspots were manually classified by comparing extracted sunspots with corresponding active region maps (ARMaps) from the Mees Observatory at the Institute for Astronomy, University of Hawaii.

Key words: sunspots, recognition, machine learning, data mining

1 Introduction

Data mining is about finding patterns in data by using computers. Finding these patterns can lead to new insights that furthers understanding about a specific domain. To take advantage of modern computers the process should be automatic or semi-automatic. The patterns discovered must be meaningful in the context of data representation, to allow for successful predictions on future data.

Machine learning is a field where the techniques for finding and describing structural patterns are developed. The word *learning* here refers to the improvement in performance. One way of defining learning in the context of machine learning is that "things learn when they change their behaviour in a way that makes them perform better in the future" The learning can then be tested by observing the behaviour and comparing with past behaviour [12].

Machine learning has been successfully applied to many real-life problems ([7], [8]) and there have been many success stories concerning the use of machine learning for classification, decision making and prediction. In *classification learning*, a learning scheme takes a set of classified examples from which it is expected to learn a way of classifying unseen examples. Classification learning is being provided with the actual outcome for each of the training examples, here this outcome is called the *class* of the example. The success of classification learning can be evaluated by trying out the concept

description that is learned on an independent set of test data for which the true classification is known but not available to the machine. The success rate on the test data provides a measure of how well the concept has been learned.

One of the practical issues with inputting data to the learning process is its preparation. Preparing input data often consumes the bulk of the effort invested in the entire data mining process [4]. The data needs to be gathered together, assembled, integrated, and cleaned up. Integrating data from many sources, as often the case in real life scenarios, present many challenges. As there may be different data formats, conventions, time periods and degree of aggregation. Sometimes one must acquire data from different domains and aggregate them together. Because so many issues are involved it is seldom possible to arrive at a satisfactory dataset the first time, giving long lead times for data preparation.

There are four important issues with input data that need to be taken into account before applying a learning scheme. These are: attribute types, missing values (not applicable), inaccurate values, and knowledge about the data [12]. It is important to scan through the data for inaccurate values and attributes. Typographical or measurement errors in numeric values generally cause outliers. Sometimes finding inaccurate values requires specialist domain knowledge. Duplicate data is another source of error since repetition will almost certainly cause learning schemes to produce different results.

There are different ways for representing patterns discovered. The output from a machine learning scheme usually takes the form of decision trees and classification rules, which are basic knowledge representation styles [12]. The word *knowledge* is used to refer to the structures that learning methods produce. There are more complex types of rules that allow exceptions to be specified and relations between attributes of different instances to be expressed. Output can also be represented using an instance-based representation or clusters.

In the remainder of this paper we will explore the above concepts and investigate the various fundamental problems of learning to classify data sets by directly investigating a practical application. The problem we have chosen is the automatic classification of sunspots by mining the vast data sets that arise in solar astronomy. This is an important physical problem, which we now introduce in the next section.

2 Sunspot observation

Sunspots have been the subject of interest to astronomers and astrophysicists for many years: sunspot sightings were first recorded in China as far back as 165 BC; Galileo made some of the first detailed hand-drawings of sunspots in 1610 using a primitive telescope. With the advent of more sophisticated telescopes and photographic devices, knowledge about sunspots and their

relationship to other solar phenomena has increased. Nowadays it is known that sunspots do not appear to be randomly scattered over the Sun's surface but are confined to a specific band. Sunspots are also recognised to have their own life-cycle. They are born and die, grow and shrink in size, form groups and formations, and move across the Sun's surface throughout their lifetime.

Sunspot observation, analysis and classification form an important part in furthering knowledge about the Sun, the solar weather and its effect on earth. Certain sunspot groups are associated with solar flares that are monitored by observatories around the world daily. These observatories capture images of the Sun's surface and make note of all the sunspots in an effort to predict solar flares. This was vividly demonstrated recently with the spectacular flares on the 23rd of October 2003 caused by a large sunspot group.

This paper describes an initial attempt at automatic sunspot recognition and classification.

3 Sunspot classification schemes

Sunspots appear on the solar disk as individual spots or as a group of spots. Larger and more developed spots have a dark interior called the *umbra*, surrounded by a lighter area referred to as *penumbra*. Sunspots have strong magnetic fields. *Bipolar* spots have both magnetic polarities present, whereas *unipolar* have only one. Within complex groups the leading spot may have one polarity and the following spots the reverse, with intermediate a mixture of both. Sunspot groups can have an infinite variety of formations and sizes, ranging from small solo spots to giant groups of spots with complex structure. Nevertheless astronomers have been able to define broad categories of sunspot groups.

Using the McIntosh Sunspot Classification Scheme [9] [10] spots are classified according to three descriptive codes.

The first code is a modification of the old Zurich scheme [3], with seven broad categories (sometimes referred to as the Modified Zurich scheme):

- A : Unipolar group with no penumbra, at start or end of spot group's life
- B : Bipolar group with penumbrae on any spots
- C : Bipolar group with penumbra on one end of group, usually surrounding largest of leader umbrae
- D : Bipolar group with penumbrae on spots at both ends of group, and with longitudinal extent less than 10 arc seconds (120 000 km)
- E : Bipolar group with penumbrae on spots at both ends of group, and with longitudinal extent between 10 and 15 arc seconds (120 000 km and 180 000 km)
- F : Bipolar group with penumbrae on spots at both ends of group, and length more than 15 arc seconds (above 180 000 km)

H : Unipolar group with penumbra. Principal spot is usually the remnant leader spot of pre-existing bipolar groups

The second code describes the penumbra of the largest spot of the group and the third code describes the compactness of the spots in the intermediate part of the group [9] [10]. Up to sixty classes of spots are covered, although not all code combinations are used. A particular spot or group of spots may go through a number of categories in their lifetime. Solar flares are usually associated with large groups.

3.1 Issues with classification

When attempting automated classification the following issues need to be taken into account:

1. Interpreting classification rules

As only broad forms of classification exist there is a large allowable margin in interpretation of classification rules. The same group may be assigned a different class depending on the expert doing the classification. Astronomical observatories around the world share information and cross-check results regularly to form an opinion. For example class E is notorious for appearing in various shapes and forms. Class E is defined in the Modified Zurich scheme (the first code from the McIntosh scheme), as a *"bipolar group with penumbrae on spots at both ends of group, and with length between 120000 km and 180000 km"*.

An interpretation of that particular classification rule might be as follows. Bipolar refers to magnetic polarities, implying that there are two or more spots in the group. Spots at both ends of group should have penumbrae. In low resolution images this may not be easy to detect. The length of a group is defined as more than 120000 km but less than 180000 km. It is however unclear between exactly which two points this should be measured. Therefore as long as there is a group with more than one spot, and an approximate spread from one end to the other between 120000 km and 180000 km in length with spots at both ends having penumbrae, the whole group could be classified as E.

2. Individual spots and groups

Sunspot classification schemes classify sunspot groups not individual spots. An individual spot may belong to any class of group. When sunspots are extracted from digital images they are treated as individual spots. Hence further information is required to group spots together to form proper sunspot groups.

There are two possible ways to arrive at the classification of a particular group. An individual spot can be assigned a class and then grouped

together with other spots of the same class. The alternative way is to group spots together first and then work out the group's class. The former method requires using additional information, for example about the spot neighbourhood, to determine its class membership and grouping it with the neighbouring spots of the same class to form proper groups. The latter method relies on image processing techniques to create boundaries around groups those regions are then treated as groups.

It can be argued that image processing techniques alone, without a priori knowledge, are not capable of reliably grouping spots, at least on a large sample of images. For example, suppose that there are two closely located sunspot groups of class F. As these groups have many spots widely spread over a horizontal space, with the two leading spots very far apart from each other there is a possibility that a region growing technique may create a single region containing spots from both groups. It is thus safer to determine an individual spot's class membership before attempting to form groups.

3. Dealing with groups migration

Sunspots start their life as small *pores*, appearing as small dark areas. In this initial stage spots are small, penumbra-less and may disappear within a few hours. Larger spots that survive this initial stage usually form pairs that may evolve into groups. These pairs usually experience a quick growth in area and separation in the first few days. As they grow their penumbra develops until maximum area and complexity is attained. The *leading spot* is usually the largest and nearly circular in outline, while several smaller spots make up the following part of the group with the smallest spots appearing in between. A pair will separate usually initially by the subdivision of the following spot into smaller spots. The leading spot retains its outline and penumbra and only gradually decreases in size. It can survive for several weeks after the disappearance of the rest of the group.

This life-cycle and migration across the solar disk have important implications for automatic recognition and classification. Firstly, a particular group will change its class assignment several times during its lifetime. A reliable method to keep track of those changes must be devised to correctly follow a group during its lifetime. It may be difficult to decide exactly when the change occurs. An individual image of a solar disk containing sunspots has no information about their previous and future class. Secondly, as groups approach the edge of the visible solar disk their shape appears compacted. This is because images of the Sun are taken from a fixed observation point. For example, a large group of class F, with dozens of spots, may still retain its actual class when approaching the edge. However because the group is at the edge of the visible solar

disk, its shape is compacted so much that it hardly resembles the class F described by classification rules.

4. Availability of data

The average number of visible sunspots varies over time, increasing and decreasing on average over 11.8 years. *Solar minimum* is the part of the cycle with low sunspot activity, whereas *solar maximum* refers to the part with high sunspot activity [3]. As each cycle progresses the sunspots gradually start to appear closer and closer to the Sun's equator.

This creates an issue when deciding on the input data range in constructing a *training dataset*. Ideally a representative sample would be chosen containing an equal amount of all classes. Choosing a suboptimal sample may result in a dataset that is biased towards several classes. The availability of certain classes during any date range must be considered. During sunspot maximum it appears that many groups of class D, E, and F are present, whereas during sunspot minimum there are more groups of class A, B, and C. It is however possible to mix and match images from different date ranges in order to balance the dataset.

5. Quality of input data

For automatic recognition and classification systems to perform they need a consistent set of high quality input images. This means providing images that are free of distortions and fairly high resolution. Images should be taken from one source and the same instrument to reduce the variability. This means satellite images are more suitable than photographs taken from the ground.

Some sunspots can be very small and may not be captured or may be filtered out by noise reduction algorithms used subsequently in image processing tools. Sunspots' physical texture makes it difficult to separate the umbra and penumbra of spots.

Given the issues above, the attempt made was to classify sunspot groups according to the seven-class Modified Zurich scheme. Individual spots were classified first before being grouped together. Satellite images from the NASA SOHO satellite's MDI¹ instrument² were used.

¹ Scherrer, P. H., et al., Sol. Phys., 162, 129, 1995.

² Daily MDI Intensitygram images at <http://sohowww.nascom.nasa.gov> - all images for this source are *Courtesy of SOHO/MDI consortium. SOHO is a project of international cooperation between ESA and NASA*

4 The design of sunspot classification system

A typical sunspot classification system consist of two modules: the image processing module and the classification module. The aim of the former is to handle the input image, extracting spots and their properties. The classification module is responsible for predicting the spot's class and grouping them together based on the information provided by the image processing module. A future objective would be to build a complete system whose input is an image and output are sunspot groups marked and classified.

In this attempt an open source image processing tool called ImageJ³ was used and modified to perform the required tasks of the image processing module. These were, taking satellite images from SOHO and performing pre-processing to remove unnecessary features leaving just the solar disk and visible spots. Next, individual spots were separated from their background using a custom threshold function and their features extracted to a text file. The process was repeated for each image so that a large flat-file was created containing all detected spots and their attributes. Such a matrix of instances and attributes was then ready to be input into machine learning tools for learning and building a classifier.

The issue of image resolution is important as the higher the resolution the more detail is available. Detail, in this context, means clear umbra and penumbra separation in spots and the finer details such as the notched appearance of spots, and presence of a multitude of pores. The images obtained from SOHO were only low resolution so no attempt was made to separate the penumbra and umbra of spots and this was left for future work.

5 Learning sunspot classification

Data mining and machine learning techniques can help to find the set of rules that govern classification and deal with the margin that exists for the interpretation of sunspot classification rules. This is achieved by learning from actual data and the past experience of expert human astronomers who have been classifying sunspots manually for years. The only prerequisite is high quality data. High quality data can be defined as a well constructed dataset compiled with the help of a domain expert. With such a dataset machine learning schemes can be applied to find strong rules; rules may then be used for classifying new, unseen, examples.

5.1 Attribute selection

Selecting the right set of attributes for use in the dataset can have a dramatic impact on the performance of the learning scheme and requires an

³ developed by Wayne Rasband, the latest version at <http://rsb.info.nih.gov/ij/download>

understanding of the problem to be solved through consulting with an expert. Limitations arise from the data source and pre-processing by the image processing module.

The features extracted by the image processing method were mostly shape descriptors describing the shape of single sunspots but containing no information about the spot's neighbours. One way of obtaining such information would be to calculate the distances to the nearest neighbours or to count how many spots are within a certain radius of the target.

For example a spot that is located somewhere inside a group of class F would be expected to have many neighbors. This can be contrasted with a spot of class H that has no immediate neighbors. Moreover within each bipolar group, there are always one or two leading spots, which are substantially larger than the rest of the spots in the group. Moving from class B to F these leading spots gets larger in size. Therefore, for any spot if the *number of neighbours*, within a certain *radius*, and their *sizes* could be determined it would almost certainly be possible to tell which class the spot belongs to.

This means that the distances between every single spot identified in an image were needed. The value of the radii used to group spots in this experiment were set to reflect 120000 km and 180000 km intervals specified in the Modified Zurich scheme. Radii were set at 60000 km, 120000 km, 180000 km. These values were converted to distances in pixels and scaled. Counts of the number of spots within each radius were computed. Future work could include the calculation of "size dimensions" of neighboring spots within a specified radius.

The following sunspot features were extracted: x and y coordinates of a spot center; *area* of a spot; *perimeter* length around a spot; spot's *angle* to the main axis; spot's *aspect ratio*, *compactness*, and *form factor*; spot's *feret's diameter*; spot's *circularity*; count of how many neighbouring spots are within a specified *radii* (nine radii were selected).

5.2 Data preparation

Another important issue is the process of dataset preparation, particularly manual classification of extracted sunspots. A reliable source of data to compare was found. The archive of ARMaps images⁴ is a collection of images that map the active regions on the sun's surface. These are regions with high sunspot counts which are clearly marked.

Issues arose when using the ARMaps images for comparison and manual classification. An example would be that the images are taken at timed intervals which may not correspond to the exact time of the NASA SOHO satellites' images.

⁴ See <http://www.solar.ifa.hawaii.edu/ARMaps/armaps.html> - the data is courtesy of *The Solar Group at the Institute for Astronomy*, Mees Solar Observatory on Haleakala, Maui.

The sunspot classification that comes with the ARMaps seems to originate from the *USAF/NOAA Solar region summary* data and it is always one day behind the date of the actual active region map displayed in the ARMap. It was therefore necessary to check one day ahead to see if the groups displayed had changed their classes. The classification used was the McIntosh scheme from which the Modified Zurich scheme was easily derived by taking the first letter of the code.

The process of constructing the *training dataset* consisted of gathering data from two sources: the NASA/SOHO website and the ARMaps pages from the Hawaii University website. Data was collected from two years worth of daily sunspot and active regions maps. Processing such an amount of data would have been very time consuming and therefore it was decided to take three months of sunspot data from September 2001 to November 2001. This gave a total of 89 satellite images and 89 active region maps.

The manual classification process was as follows and was repeated for all 89 images. Found an ARMap that fitted the corresponding drawing of detected sunspots using the date and the filename of a drawing. Looked at the regions marked on the ARMap and matched them with the regions of spots detected in the drawings. All regions on the ARMap were numbered - to be annotated. All spots that fell within each identified region were selected. Since each spot is numbered, it was possible to assign the ARMap region number to those spots in the main flat file. All spots with an identical ARMap region number were assigned the class of the ARMap region.

One issue concerned the number of spots detected by the image processing system. Some of the finer details were not detected, largely due to resolution issues. It was discovered that occasionally groups that were classified as class B or C in the ARMaps could only be classified as H from the drawings. These bipolar classes have one leading spot and several very tiny following spots. In the drawings these following spots were not detected meaning the whole group could only be classified as class H rather than B or C. However, it would be dangerous to treat all these spots as class H. Usually class H spots are very large, single spots with no neighbours. Therefore in the end it was decided to mark these spots as H if they were of a sufficient size. Otherwise they were left out altogether.

As sunspot groups change their shape and become smaller in size as they approach the edge of the solar disk there is an increased possibility of misclassification. For example a large group of class F approaching the edge of the solar disk, would still be classified in the ARMap's region summary data as class F, while it had only two or three visible spots. Normally such a group would have been classified as class B or C at best. Where applicable this has been dealt with by not taking those groups into account.

In summary a total 2732 examples were manually classified, of which 143 were either those that were left out due to the issues explained above or misidentified spots. "Misidentified" means those spots that have been recog-

nized in the drawings but not classified on the ARMaps or vice-versa. Overall there were 2589 instances giving a *misidentification rate* of 5.23%.

5.3 Learning methods

Two data mining tools WEKA [12,14] and RSES[13,1] were used. They contain all the learning schemes implemented. The classification "success rate" was determined by the number of *true positives* and *true negatives* over the entire range of classes. This meant that on the resulting confusion matrix high values across the main diagonal line should have been seen.

6 The results of experiments

We performed two series of experiments with classification algorithms. In the first series, we applied four well-known classification algorithms on the prepared data set (containing 2589 objects and 20 attributes), namely:

WEKA.J48: The implementation of C4.5[11] decision tree algorithm in WEKA system.

WEKA.Ibk: The implementation of kNN algorithm in WEKA system.

RSES.LEM2: The implementation of LEM2[6] algorithm in RSES system.

RSES.kNN: The RIONA algorithm[5] – the classification algorithm combining rule induction and instance based learning methods. This method is implemented in RSES system.

The distribution of classes in our data set is presented in the Table 1

Group classification	A	B	C	D	E	F	H
Class distribution	0.31%	1.62%	7.49%	30.67%	25.45%	28.51%	5.95%

Table 1. The distribution of classes in the dataset

In the second series of experiments, before applying previous classification methods, we selected the most relevant subset of attributes for each learning algorithm. For most algorithms the best subset consisted of attributes describing spots neighbourhood and location. Shape descriptors were less relevant. In addition a boosting method, called the *AdaBoostM1* [2], was applied to the J48 algorithm to improve results. Experiment results are presented in Table 2.

7 Conclusions

We have conclusively demonstrated that the automatic classification of sunspots is possible and the results show that a high degree of accuracy can be

Scheme	Accuracy	A	B	C	D	E	F	H
J48 all attributes	73.31 %	0.13	0.33	0.54	0.73	0.73	0.80	0.84
J48 subset	77.33 %	0	0.36	0.60	0.80	0.77	0.83	0.80
J48 subset + boost	85.09 %	0	0.57	0.72	0.88	0.86	0.88	0.81
IBk all attributes	63.89 %	0.25	0.29	0.45	0.66	0.65	0.71	0.54
IBk subset	89.57 %	0.25	0.76	0.85	0.92	0.91	0.94	0.62
RSES kNN all	83.32 %	0.20	0.65	0.72	0.84	0.85	0.86	0.84
RSES kNN subset	90.60 %	0.13	0.59	0.79	0.91	0.94	0.94	0.78
RSES LEM2 all	66.84 %	0.10	0.47	0.46	0.65	0.68	0.72	0.84
RSES LEM2 subset	77.50 %	0	0.55	0.58	0.79	0.80	0.81	0.77

Table 2. Comparison of accuracy and true positive rates of different classification algorithms

achieved. High classification accuracy was obtained for sunspot groups D, E, F, and H. For groups A, B, and C only low classification accuracy was achieved due to the distribution of classes in the training dataset (see Table 1. In future work we are planning to improve the image processing module to extract additional attributes and enriching the training dataset with new examples. These changes should improve the accuracy of classification. We are also planning to apply clustering algorithms to build a multi-layered classifier and extending it to cover the entire McIntosh scheme. The ultimate goal is to build a complete sunspot classification system.

References

1. Bazan J., Szcuzka M. RSES and RSESLib - A Collection of Tools for Rough Set Computations, Proc. of RSCTC'2000, LNAI 2005, Springer Verlag, Berlin, 2001
2. Freund, Y., and R. E. Schapire *Experiments with a new boosting algorithm*. Proc. Thirteenth International Conference on Machine Learning, Morgan Kaufmann, 1996, pages 148-156.
3. R. J. Bray and R. E. Loughhead. *Sunspots*. Dover Publications, New York, 1964.
4. P. Hadian R. Stadler J. Verhees Cabena, P. and A. Zanasi. *Discovering data mining: From concept to implementation*. Prentice Hall, Upper Saddle River, NJ., 1998.
5. G. Gora and A. Wojna., RIONA: A New Classification System Combining Rule Induction and Instance-Based Learning, *Fundamenta Informaticae*, 51(4), 2002, pages 369-390
6. Grzymala-Busse J., A New Version of the Rule Induction System LERS *Fundamenta Informaticae*, Vol. 31(1), 1997, pp. 27-39
7. R. Kohavi and F. Provost. Machine learning: Special issue on application of machine learning and the knowledge discovery process. *Machine Learning*, 30, 1998.
8. P. Langley and H. A. Simon. Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):55-64, 1995.
9. K. J. H. Phillips. *Guide to the Sun*. Cambridge University Press, 1992.

10. P. McIntosh, Solar Physics 125, 251, 1990.
11. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
12. I. H. Witten and Frank E. *Data Mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, San Francisco, CA., 2000.
13. The RSES Homepage, <http://logic.mimuw.edu.pl/~rses>
14. The WEKA Homepage, <http://www.cs.waikato.ac.nz>