Rough Set Approach to Sunspot Classification Problem

Sinh Hoa Nguyen[1], Trung Thanh Nguyen[2], and Hung Son Nguyen[3]

[1] Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008, Warsaw, Poland
[2] Department of Computer Science, University of Bath,
Bath BA2 7AY, United Kingdom
[3] Institute of Mathematics, Warsaw University,
Banacha 2, 02-097 Warsaw, Poland

Abstract. This paper presents an application of hierarchical learning method based rough set theory to the problem of sunspot classification from satellite images. The Modified Zurich classification scheme [3] is defined by a set of rules containing many complicated and unprecise concepts, which cannot be determined directly from solar images. The idea is to represent the domain knowledge by an ontology of concepts – a treelike structure that describes the relationship between the target concepts, intermediate concepts and attributes. We show that such ontology can be constructed by a decision tree algorithm and demonstrate the proposed method on the data set containing sunspot extracted from satellite images of solar disk.

Keywords: Hierarchical learning, rough sets, sunspot classification.

1 Introduction

Sunspots that appear as dark spots on the solar surface, have been the subject of interest to astronomers and astrophysicists for many years. Sunspot observation, analysis and classification form an important part in furthering knowledge about the Sun, the solar weather, and its effect on earth [8]. Certain categories of sunspot groups are associated with solar flares. Observatories around the world track all visible sunspots in an effort to early detect flares. Sunspot recognition and classification are currently manual and labor intensive processes which could be automated if successfully learned by a machine.

Some initial attempts at automatic sunspot recognition and classification were presented in [4]. Several learning algorithms were examined to investigate the ability of machine learning in dealing with the problem of sunspot classification. The experiment showed that it is very difficult to learn the classification scheme using only visual properties as attributes. The main issue is that many characteristics of sunspots can not be precisely determined from digital images.

To improve the classification accuracy we experimented with classification learning in combination with clustering and layered learning methods. It was

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3642, pp. 263–272, 2005.

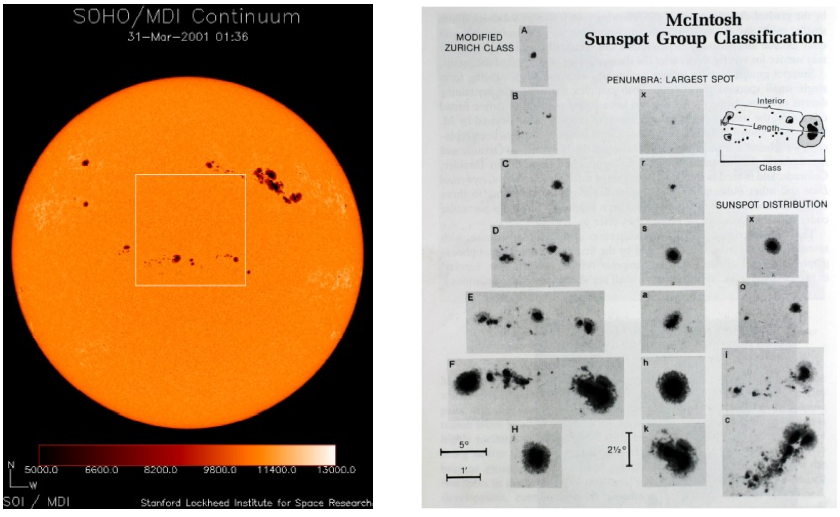

Fig. 1. Left: The SOHO/MDI satellite image of the solar disk, showing sunspots. Right: the McIntosh Sunspot Classification Scheme. (Courtesy P.S. McIntosh, NOAA(1990).

concluded that one possible way of improving accuracy is to embed the domain knowledge into the learning process. In previous papers we have considered the case where domain knowledge was given in a form of concept ontology and have presented a rough set and layered learning based method that successfully makes use of such kind of domain knowledge [5] [7]. In this paper, that approach is applied to the sunspot classification problem with an exception that the concept ontology is *not given* but constructed by a supervised learning method. The proposed solution has been implemented and the experimental results show many advantages in comparison with standard learning algorithms.

2 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. Despite such a diversity of shapes and sizes astronomers have been able to define broad categories of sunspot groups. Using the McIntosh Sunspot Classification Scheme [3] spots are classified according to three descriptive codes. The first code is a modification of the old Zurich scheme, with seven broad categories:

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
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 [3]. 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. When attempting automated classification the following issues need to be considered:

1. **Interpreting classification rules:** As only broad forms of classification exist there is a large allowable margin in the interpretation of classification rules. The same group may be assigned a different class depending on the expert doing the classification. Observatories share information and cross-check results regularly to form an opinion.
2. **Individual spots and groups:** Sunspot classification schemes classify sunspot groups not individual spots. 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.
3. **Dealing with groups migration:** Sunspots have their own life-cycle and migrate across the Sun's surface. They start their life as small tiny spots that usually continue to form pairs and evolve into groups. Once a group attains its maximum size it starts to decay. As a result, a particular group may 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. Moreover, as groups approach the edge of the visible solar disk their shape appears compacted making classification based solely on digital images difficult.
4. **Availability of data:** The average number of visible sunspots varies over time, increasing and decreasing on average over 11.8 years. As each cycle progresses sunspots gradually start to appear closer to the Sun's equator while forming larger and more complex groups. This creates an issue when deciding on the input data range for a *training dataset*. For example by

taking observations only from a short period at solar maximum, where there are likely to be more sunspots groups class D, E, F, an unbalanced training sample may be obtained.

5. **Quality of input data:** For automatic recognition and classification systems to perform well they need a consistent set of high quality input images, free of distortions and of fairly high resolution. Images should be taken from one source and the same instrument to reduce the variability. Thus satellite images are more suitable than photographs taken from the ground. Note that some sunspots can be very small and may not be captured at all.

The automated sunspot classification system that we propose 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.

Our current system is able to import digital images of solar disks from NASA SOHO/MDI satellite, separate individual spots from their background using a custom threshold function and extract their features to a text file to build a matrix of instances and attributes. Such a flat-file can be imported to machine learning tools (such as WEKA, RSES) for building a classifier. A future objective would be to build a complete system whose input is an image and output are sunspot groups marked and classified.

3 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 standard learning algorithms that used only visual properties to predict classification scheme proved to be inadequate, especially for robust and accurate daily prediction. To improve the classification accuracy, it is necessary to embed the domain knowledge into the learning process. This paper presents a learning method to sunspot classification based on rough sets and layered learning approach. Layered learning [11] is an alternative approach to concept approximation. Given a hierarchical concept decomposition, the main idea is to synthesize a target concept gradually from simpler ones. One can imagine the decomposition hierarchy as a treelike structure containing the target concept in the root. A learning process is performed through the hierarchy, from leaves to the root, layer by layer. At the lowest layer, basic concepts are approximated using feature values available from a data set. At the next layer more complex concepts are synthesized from basic concepts. This process is repeated for successive layers until the target concept is achieved. In previous papers (see [6] [5]) we presented a hierarchical learning approach to concept approximation based

on rough set theory. The proposition was performed with an assumption that the concept ontology already exists. This assumption is not satisfied in the case of sunspot classification problem. One of the main issues of this contribution is the construction of concept ontology from the domain knowledge. Our solution to sunspot classification problem consists of four main steps:

1. recognize single sunspots using image processing techniques and create decision table describing their classification made by experts;
2. group daily sunspots into clusters and create decision table for those clusters;
3. create a concept ontology from the domain knowledge;
4. apply hierarchical learning method based on rough set theory to learn the Zurich sunspot classification scheme.

3.1 Sunspot Recognition and Data Preparation

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. The method of sunspot recognition and extraction from digital images of solar disk was described in [4]. The resulting data set consists of sunspots as objects, their visual properties (size, shape, etc.) as attributes and the Zurich classification (made by experts from ARMaps) as the class label.

Attribute Selection: The features extracted by the image processing method were shape descriptors describing the shape of single sunspots and information about spot's neighbours. 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).

Data Preparation: The following manual classification process was repeated for all training 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.

3.2 Sunspot Clustering

For each image, individual spots were grouped together using a simple hierarchical agglomerative clustering algorithm. The objective was to obtain groups which closely matched real life sunspot groups. Three different methods were used and compared: single-link, complete-link and group average [2]. The Euclidean distance was used to calculate the dissimilarity measure. The clustering process

starts with all spots within a single image. Spots are then merged into groups until the stop condition is triggered. The stop condition was based on the total distance of all spots within a single cluster. If at any iteration that total distance across all clusters exceeded a predefined threshold then the process is stopped and groups produced.

3.3 Construction of the Concept Ontology

The main goal of sunspot classification problem is to classify recognised sunspots into one of the seven classes $\{A, B, C, D, E, F, H\}$. After the clustering step, the task is restricted to classification of sunspot groups. In our system, every cluster is characterised by about 40 attributes. These attributes describe not only properties of whole groups, but also features of the largest spots in a group.

In Section 1 we have presented the original sunspot classification scheme. This scheme seems to be complicated but, in fact, the classification can be described by some simpler concepts:

1. **Magnetic type of groups:** there are two possible types called *unipolar* and *bipolar*;
2. **Group span:** a heliographical distance of two farthest spots in a group; there are three spanning degrees, i.e., *NULL* (not applicable), *small* (less than 10 h.degs. or 120000 km), *large* (more than 15 h.degs. or 180000 km) and *middle* (between 10 h.degs and 15 h.degs.);
3. **Penumbra type of the leading spot:** there are four possible types called *no penumbra, rudimentary, asymmetric, and symmetric*;
4. **Penumbra size of the leading spot:** there are two possible values *small* (less than 2,5 h.degs. or 30000km), and *large* (more than 2,5 h.degs.);
5. **Distribution of spots inside a group:** there are four possible values called *single, open, intermediate, and compact.*

If we consider all situations described by those five concepts, one can see that there are 60 possible situations only. Every situation is characterized by those concepts (which can be treated as attributes) and can be labeled by one of seven letters $\{A, B, C, D, E, F, H\}$, accordingly to the Zurich classification scheme. Therefore we have a decision table with 60 objects, 5 attributes, 7 decision classes. The idea is to create a decision tree for the described above decision table. The resulting tree computed by the decision tree induction method, which is implemented in Weka [14] as J48 classifier, is presented in Figure 2.

This decision tree leads the following observations, which are very useful for concept decomposition process: (1) Classes D, E and F are similar on almost all attributes except attribute **group span**; (2) Classes A, H have similar magnetic type (both are unipolar), but they are discerned by the attribute **penumbra type**; (3) Classes B, C have similar magnetic type (both are bipolar), but they are discerned by the attribute *penumbra size*.

The final concept ontology of target concept has been build from those observations. Figure 3 presents the main part of this ontology which was created by including the following additional concepts to the decision tree in Fig. 2:

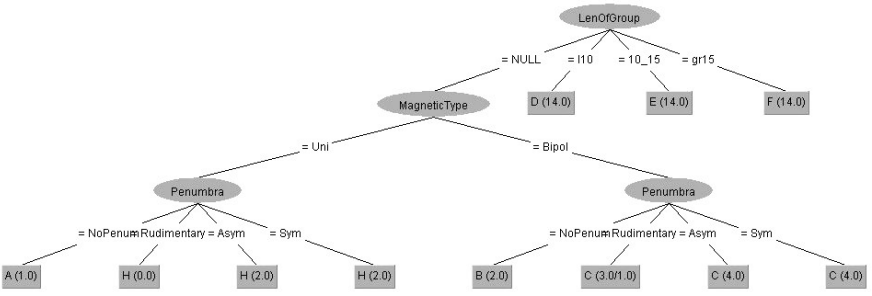

Fig. 2. The Zurich classification scheme represented by a decision tree

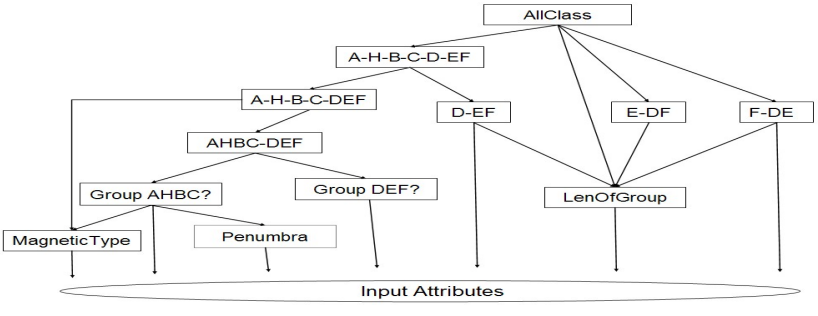

Fig. 3. The concept hierarchy for sunspot classification problem

- **Group AHBC?**: does a sunspot cluster belong to one of classes A, B, C, H?
- **Group DEF?**: does a sunspot cluster belong to one of classes D, E, F?
- **AHBC-DEF**: the classification distinguishing $\{A, B, C, H\}$ and $\{D, E, F\}$;
- **A-H-B-C-DEF**: the classification that groups classes D, E, F together;
- **A-H-B-C-D-EF**: the classification that groups classes E, F together;
- **D-EF**, **E-DF**, **F-DE**: classification problems that distinguish one class from the rest for three decision classes D, E, F;
- **target classes**: what is the label of a sunspot cluster?

The synthesis process is performed through the concept hierarchy, from leaves to the root as it has been presented in [5]. The learning algorithm, for every node N of the concept hierarchy, produces the rough membership function for every decision class that occurs in N. Later, the extracted membership functions are used as attributes to construct the rough membership function for those concepts occurring in the next level of the hierarchy.

We have shown that rough membership function can be induced by many classifiers, e.g., k-NN, decision tree or decision rule set. The problem is to chose the proper type of classifiers for every node of the hierarchy. In experiments with sunspot classification, we have applied the rule based classification algorithm and the modified nearest neighbor algorithms that were implemented in RSES [13].

4 The Results of Experiments

In previous paper, we have performed some experiments with classification of single sunspots. The prepared data set contains 2589 sun spots (objects) extracted from 89 daily images of solar disk (from Sep 2001 to Nov 2001). Each object was described by 20 attributes and labeled by one of the decision class A, B, C, D, E, F, H.

In this paper we consider a temporal testing model where the training data set contains those spots that occur within first two months, i.e., from Sep. 2001 to Oct. 2001, and the testing data set contains those spots that occur in the last month, i.e., in Nov. 2001. Classification accuracies of standard learning algorithms for such data sets are very poor and oscillate about 38%. Applying the proposed method one can improve the classification accuracy.

4.1 Sunspot Clustering

Because most real life sunspot groups are either compact or elongated it was difficult to choose between the single-link and complete-link method. Complete-link method produced more compact clusters but failed to uncover elongated groups correctly. Single-link method, on the other hand, suffered from clustering too many distinct groups together. The group average method was also used but the results obtained were not as good as the complete-link method, which proved to be the best compromise for the given data. It produced many compact but correct groups contained within larger elongated groups instead of small number of large but incorrect elongated groups.

Since sunspot groups have dimension limits the sum of all spot distances within a cluster was used for a stopping condition. If a diameter of a cluster grows too large the clustering process is stopped. The experiments were made to obtain the best threshold value. A performance measure used for obtaining the best threshold value was a cluster purity measure. For each cluster produced by the clustering algorithm a comparison was made with the reference cluster to identify how many spots were in fact correctly grouped. A 100% pure cluster is the cluster which had all the spots correctly grouped. So to find the best threshold value for the dataset the cluster purity measure was calculated for each cluster and the average obtained for the whole dataset for every threshold value. The threshold value which produced the best average was ultimately chosen.

4.2 Classification of Sunspot Clusters

For each daily image of solar disk in the three month period from September 2001 to November 2001, we have applied the sunspot recognition algorithm and the described above clustering algorithm to extracted sunspots. Total of 494 sunspot clusters were obtained. The train set (obtained from September and October 2001) consists of 366 clusters, while the test set (November 2001) contains 128 sunspot clusters. The distribution of decision classes in training and testing data is presented in Table 1.

Table 1. The distribution of decision classes on training and test data sets

Train/Test table	No of obj.	Zurich's classes						
		A	B	C	D	E	F	H
Train set	366	0,8%	2,2%	9,6%	30,6%	19,7%	21,9%	15,3%
Test set	128	0%	1,6%	7,8%	36,7%	18,8%	18%	17,2%

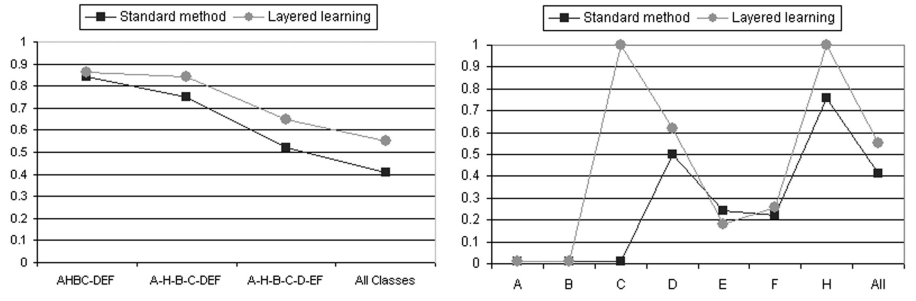

Fig. 4. Left: The classification accuracy of standard and layered method for some concepts in the ontology presented in Fig. 3. Right: the classification accuracy of standard and layered method for particular decision classes.

We have performed some experiments with learning the sunspot classification to compare accuracy of two methods, i.e., the standard rule based method and the proposed method based on layered learning idea. Experimental results are shown in Figure 4. A considerable improvement was obtained by applying the proposed method based on rough sets and layered learning approach compared to standard methods. The highest improvements were achieved for classes C and H that were recognized by the layered learning method with 100% accuracy, see Fig. 4 (right). Classes A and B were too small to be evaluated. Also, the accuracy of the recognition problem: "whether a cluster belongs to one of three classes D, E, F" was very high (about 98%). The main problem here was how to separate those three classes. The decision tree presented in Fig. 2 suggests that these classes can be separated by the cluster span. Unfortunately, our clustering algorithm tends to form smaller groups compared to the real ones. Therefore some large clusters may have been divided into a few smaller ones, and this could have been the reason for low classification accuracy of classes D, E, F.

5 Conclusions

We have demonstrated that automated classification of sunspots is possible and the results show that higher accuracy can be achieved through a layered learning approach and sunspot clustering. 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 help to improve the

accuracy of classification further and address some of the shortcomings in the current training data. We are also planing to improve clustering algorithms to increase the classification quality of three classes D, E, F.

Acknowledgement. The research has been partially supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland and the research grant of Polish-Japanese Institute of Information Technology.

References

1. Bazan J., Szczuka M. RSES and RSESlib - A Collection of Tools for Rough Set Computations, Proc. of RSCTC'2000, LNAI 2005, Springer Verlag, Berlin, 2001
2. Jain, A.K., Murty M.N., and Flynn P.J. (1999): Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3, 264-323
3. P. McIntosh, Solar Physics 125, 251, 1990.
4. Trung Thanh Nguyen, Claire P. Willis, Derek J. Paddon, and Hung Son Nguyen. On learning of sunspot classification. In Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors, *Intelligent Information Systems, Proceedings of IIPWM'04, May 17-20, 2004, Zakopane, Poland*, Advances in Soft Computing, pages 59–68. Springer, 2004.
5. Sinh Hoa Nguyen, Jan Bazan, Andrzej Skowron, and Hung Son Nguyen. Layered learning for concept synthesis. In Jim F. Peters, Andrzej Skowron, Jerzy W. Grzymala-Busse, Bozena Kostek, Roman W. Swiniarski, and Marcin S. Szczuka, editors, *Transactions on Rough Sets I*, volume LNCS 3100 of *Lecture Notes on Computer Science*, pages 187–208. Springer, 2004.
6. Sinh Hoa Nguyen and Hung Son Nguyen. Rough set approach to approximation of concepts from taxonomy. In *Proceedings of Knowledge Discovery and Ontologies Workshop (KDO-04) at ECML/PKDD 2004, September 24, 2004, Pisa, Italy*, 2004.
7. Sinh Hoa Nguyen and Hung Son Nguyen. Learning concept approximation from uncertain decision tables. In Monitoring, Security, and Rescue Techniques in Multiagent Systems Dunin-Keplicz, B.; Jankowski, A.; Skowron, A.; Szczuka, M. (Eds.), Advances in Soft Computing, Springer-Verlag 2005, page 249–260.
8. K. J. H. Phillips. *Guide to the Sun.* Cambridge University Press, 1992.
9. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
10. P. H. Scherrer, et al., Sol. Phys., 162, 129, 1995.
11. P. Stone. *Layered Learning in Multi-Agent Systems: A Winning Approach to Robotic Soccer.* The MIT Press, Cambridge, MA, 2000.
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`