A Hybrid System for Learning Sunspot Recognition and Classification

Trung Thanh Nguyen, Claire P. Willis, Derek J. Paddon
Computer Science Department, University of Bath
Bath BA2 7AY, United Kingdom

Hung Son Nguyen
Institute of Mathematics, Warsaw University
ul. Banacha 2, 02-097 Warszawa, Poland
contact person email: csmttn@bath.ac.uk

Abstract

Sunspots observation and classification is an important task for solar astronomers. The activity of sunspots can give clues to the timing of solar flares and the solar weather in general. This paper describes a hybrid system for automatic sunspot recognition and classification. The system uses a combination of image processing and machine learning techniques to process and classify sunspot groups from digital satellite images. Sunspot data are extracted from daily images of the solar disk captured by the NASA SOHO/MDI satellite. The classification scheme attempted was the seven-class Modified Zurich scheme. The main components of the hybrid system are: 1) the image processor, 2) the feature extractor, 3) the clusterer, 4) the classification learner.

Furthermore, the paper compares two clustering algorithms: hierarchical average-link and a density-based DBSCAN and examines their usefulness in dealing with sunspot data. The aim is to create clusters that closely match "natural" sunspot groups. Clustering is an important step in the process as the classification performance is dependent on it.

1. Introduction

Sunspots are the subject of interest to many astronomers and solar physicists. Sunspot observation, analysis and classification form an important part of furthering the knowledge about the Sun. Sunspot classification is a manual and labor intensive process as each spot needs to be recognized, marked, and classified. Daily images of the solar disc containing sunspots are closely examined by astronomers for clues about the state of the Sun, possibility of solar flares and how solar activity can affect the weather on Earth [14]. The entire process can be improved if sunspot classification could be learned by a machine. The difficulty with automatic classification is with the non-deterministic nature of the classification.

The image data is gathered from NASA SOHO satellite in the form of daily solar disc snapshots. The data is then fed into the hybrid system where individual sunspots are extracted using image processing techniques. A table of all spots and their attributes is built for further analysis. A series of experiments were performed on the training dataset with an aim of learning sunspot classification and improving prediction accuracy. The experiments involved using decision trees, rough sets, hierarchical clustering and layered learning methods. Results have shown that it is possible to accurately classify individual sunspots using decision trees and rough sets [10]. The results can be further improved if spots are clustered and layered learning method employed [8] [7].

Three areas where improvements could be made were: (1) the training dataset was small and unbalanced thus affecting classification accuracy of several sunspot classes; (2) the clustering algorithm could be made to produce sunspot groups that more closely matched real sunspot groups (ie. fewer groups and purer); (3) the image processing module could be improved to extract more detail from input images. The new contribution of this paper is the analysis of various clustering methods and its effect on classification accuracy.

2. Sunspot classification

Sunspots appear on the solar disk as individual spots or as a group of spots. Larger and more developed spots have a dark interior, the *umbra*, surrounded by a lighter area, the *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 with complex structure [14]. Despite such a diversity of shape and sizes astronomers have been able to define broad categories of sunspot groups.

Using the McIntosh Sunspot Classification Scheme [1] 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: A bipolar group with no penumbra (no limit to the extent of the group)

C: A bipolar group with penumbra on spots of one polarity, usually on spots at only one end of an elongated group. Class C groups become compact class D when the penumbra exceeds 5degrees in longitudinal extent. There is no upper limit to the length of class C groups

D: A bipolar group with penumbra on spots of both polarities, usually on spots at both ends of an elongated group. The length does not exceed 10 degrees of heliographic longitude

E: A bipolar group with penumbra on spots of both polarities and with a length between 10 and 15 heliographic degrees

F: A bipolar group with penumbra on spots of both polarities and with a length exceeding 15 heliographic degrees

H: A unipolar group with penumbra. Attendant spots are less than 3 heliographic degrees from the penumbra of the main spot. The principal spots are nearly always the leader spots remaining from an old bipolar group. Class H groups become compact class D when the penumbra exceeds 5 degrees in longitudinal extent

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. 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.

2.1. Manual counting and classification

An expert astronomer would go through the following process to determine sunspot group classification [19].

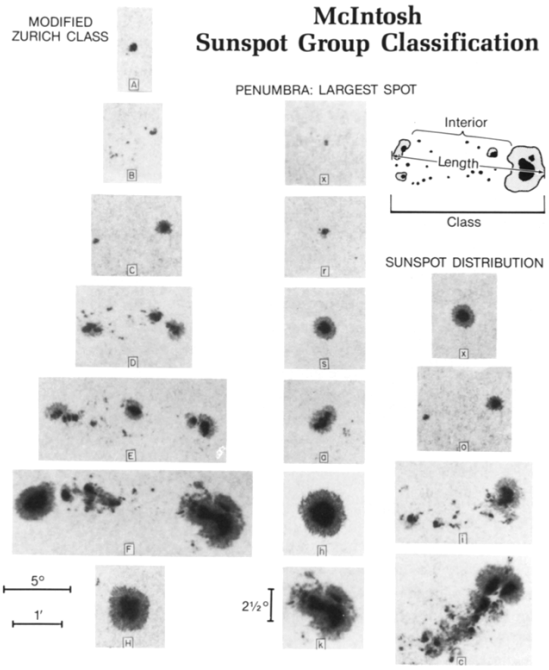

Figure 1. The McIntosh Sunspot Group Classification Scheme. (Courtesy Patrick S. McIntosh, NOAA(1990))

Starting from the northern hemisphere, west to east, then southern hemisphere, west to east, determine:

1. polarity of spots
2. the Modified Zurich letter (A or H for unipolar, B, C, D, E, or F for unipolar)
3. the second McIntosh letter:
   - is it a single spot? (x)
   - is the leader spot a rudimentary spot? (r)
   - does the leader spot have penumbra?
   - if a penumbra, is it asymmetric or is it symmetric? (a, s, k, h)
   - is the leader spot less than 2" heliographic? (a, s)
   - is the leader spot greater than 2" heliographic? (k, h)
4. the third McIntosh-letter
   - is it a single spot? (x)
   - is the spotted region open, i.e. no spots between leading and preceding spots? (o)

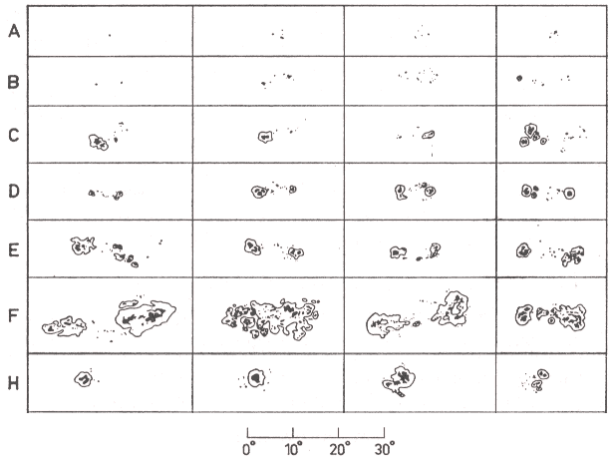

Figure 2. Allowable sunspot groups variation within a single Modified Zurich class

- is the spotted region intermediate, i.e. spots between leading and preceding spots? (i)
- is the spotted region compact, i.e. penumbra spots between leading and preceding spots? (c)

Our findings indicate that the above process is not always trivial for even a trained astronomer. The Figure 2 clearly illustrates the allowable margin for an interpretation of each class. The hybrid system described in this paper focus only on the Modified Zurich class with a plan to extend to full 60-class McIntosh scheme in the future.

3. Hybrid system design

Daily sunspot recognition and classification is a time consuming and labour intensive process. It involves many people from astronomers, computer scientists to military personnel (according to the NOAA/USAF procedures manual). The aim of the hybrid system is to aid astronomers in this process by combining image processing and machine learning techniques in one system. In order for it to be successful it needs to be robust in determining correct sunspot groups and their classes while providing an ability to accept and learn from user feedback. The design and various processing stages of the system is presented in Figure 3. The system consists of the following main modules: 1) the image processor, 2) the clusterer, 3) the repository of the domain knowledge data, 4) the classification learner.

The image processor is responsible for loading and processing of digital images of the solar disk. It extracts individual sunspots as single particles with various attributes describing sunspots' shape, size, and information about its neighbourhood [10]. The level of detail, resolution and complexity of the image processing engine determines the quality and the amount of data that can be extracted. For sunspots classification it is important to distinguish between spots with and without the penumbra and the shape and size of penumbrae. Only high resolution images can provide enough detail to successfully describe spot's penumbrae. Similarly different background subtraction functions are used depending on the level of detail (simple threshold function or an advanced segmentation algorithm).

Extracted sunspot features are stored in a table with each observation in a single row and various attributes in columns. Each spot is treated as an individual observation. Thus to recreate sunspot groups those spots need to be grouped by a clustering algorithm (more in Section 4). Clustered spots are then described as a group and provided as input for classification learners [7] [8].

The domain knowledge repository contains reference classification data and models. One example of such data is the reference sunspot groups data with classifications done by expert astronomers. There are two main information sources: 1) the daily joint NOAA/USAF Active Region Summary, and 2) ARMAps from Mees Observatory at the Institute for Astronomy, University of Hawaii. The collated information permits manual matching or every extracted spot with the real sunspot group, thus allowing to add classification labels to individual spots. The repository currently acts as the main source of training data for classification learning. In the future user feedback will hopefully be gathered here to help in improving classifiers.

Finally the classification module acts as both a classification learner and a classifier for new sunspot groups examples. Several popular classification algorithms are employed, including decision trees (J48 and kNN from WEKA [18], Lem2 [16] and RIONA [15] from RSES system [17] [9]) and layered learning methods [13] [11] [12].

4. Clustering

Clustering is an important stage in the hybrid system. When sunspots are extracted from digital images they are treated as single particles thus no information is available regarding their group membership. Sunspots need to be divided into groups first before the classification can take place. The challenge is to try to recreate natural clusters as discovered by a human astronomer.

4.1. Clustering algorithms

There are two basic types of clustering algorithms [4]: partitioning and hierarchical algorithms. Partitioning algorithms create a partition of a database D of n objects into a set of k clusters, where k is referred to as an input parameter

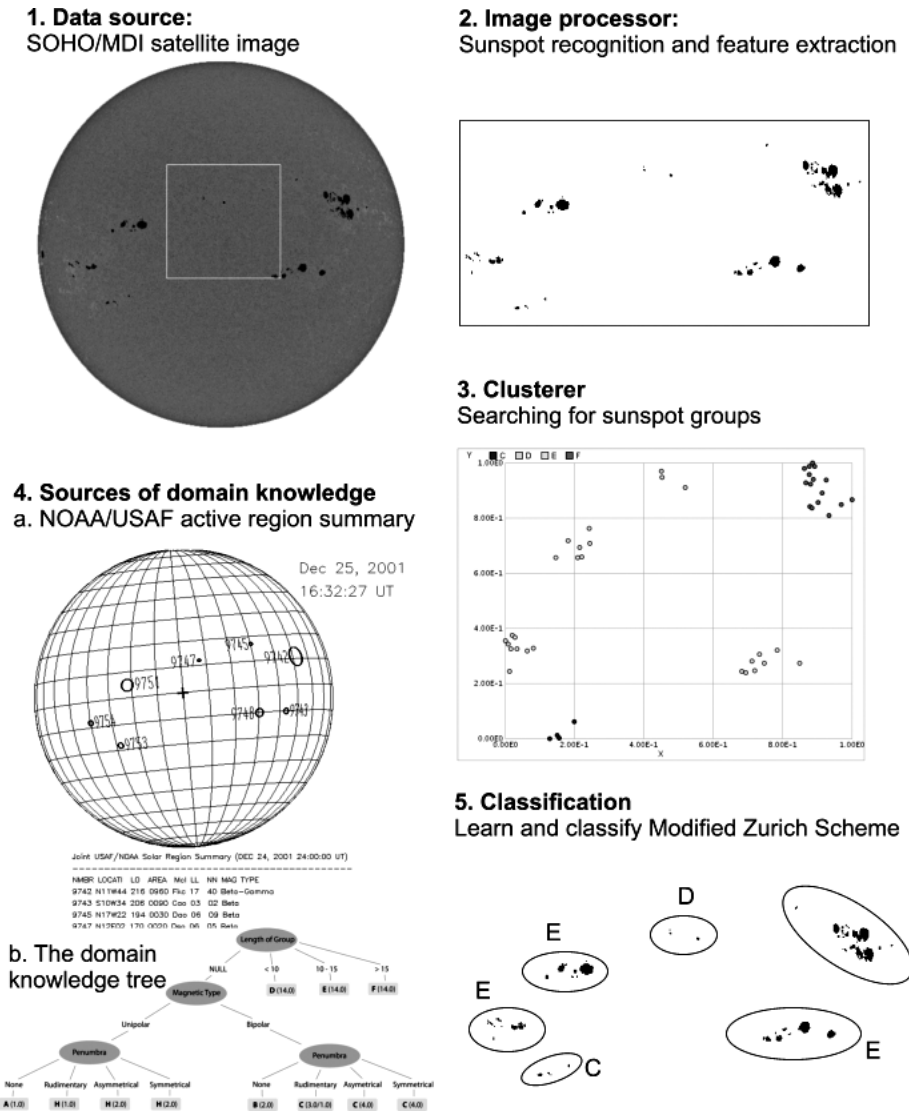

Figure 3. The architecture of the sunspot classification system

for these algorithms. However to estimate k some domain knowledge is required which is not available for many applications. The partitioning algorithm usually starts with an initial partition of D and then uses an iterative control strategy to optimize an objective function. Each cluster is represented by the center of gravity of the cluster (k-means algorithms) or by one of the objects of the cluster located near its center (k-medoid algorithms).

Hierarchical algorithms create a hierarchical decomposition of D which can be represented by a dendogram (ie. a tree that iteratively splits D into smaller subsets until each subset consists of only one object. With a such hierarchy, each node of the tree represents a cluster of D. There are two ways of creating the dendogram: from the leaves up to the root (agglomerative approach) or from the root down to leaves (divisive approach) by merging or dividing clusters at each step. Hierarchical algorithms, unlike partitioning algorithms, do not need k as an input parameter. However, a termination condition has to be defined to indicate when the merge or division process should be terminated.

Both types of algorithms have known limitations. Partitioning algorithms are effectively two stages processes. First stage involves determining k representatives minimizing the objective function. In the second stage the algorithm assigns each object to the cluster with its representative closest to the considered object. Thus a partition can be said to be equivalent to a voronoi diagram and each cluster is contained in one of the voronoi cells [4]. Therefore, the shape of all clusters found by a partitioning algorithm is convex which is very restrictive. The main issue with hierarchical clustering algorithms has been the difficulty of deriving appropriate parameters for the termination condi-

tion. It has to be a value which is small enough to separate all ”natural” clusters but large enough such that no cluster is split into two parts.

We attempted clustering individual sunspots [10] and as part of the layered learning experiment [7]. Several clustering methods were examined, including three types of hierarchical algorithms and a simple k-means algorithm. The difficulty with the latter was to find a suitable input parameter k since it is not known a priori how many sunspot groups would be present on the solar disk on a given day. Thus providing an arbitrary input parameter for the k-means algorithm would likely result in poor clusters. Finding an appropriate termination condition for hierarchical algorithms also proved difficult [8]. 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. This technique worked quite well except for large sunspot groups of elongated shapes. The hierarchical algorithm failed to detect them as single large clusters and ended up splitting them into smaller clusters thus having a negative effect on the overall classification performance of the system.

Jain [4] explores a density based approach to identify clusters in k-dimensional space. The data set is partitioned into a number of non-overlapping cells and histograms are created. Cells with high frequency counts of points are the potential cluster centers and the boundaries between clusters fall in the ”valleys” of the histogram. This method has the capability of identifying clusters of any shape.

4.2. Experiments

Encouraged by many benefits of density-based clustering a set of experiments were performed to test this technique on sunspot data. A popular and robust algorithm, called DBSCAN [**?**], was selected. Density-based clustering is based on density (local cluster criterion), such as density-connected points. Each cluster has a considerable higher density of points than outside of the cluster. The DBSCAN algorithm is as follows: 1) arbitrary select a point p, 2) retrieve all points density-reachable from p with the regards to two input parameters Eps and MinPts, 3)if p is a core point, a cluster is formed; if p is a border point, no points are density-reachable from p and the algorithm selects the next point of the database, 4) continue the process until all of the points have been processed. Eps is defined as the maximum radius of the neighbourhood, and MinPts is the minimum number of points in an Eps-neighbourhood of that point.

For prototyping purposes two set of experiments were performed on a selected sunspot dataset from August 2001 to December 2001. The first set included using hierarchical average-link clustering with a termination condition previously described in [7]. The second set included processing the same dataset with a DBSCAN algorithm with input parameters Eps = 0.1 and MinPts = 1. To estimate the input parameters some heuristic was used. Individual sunspots were described by their X and Y coordinates and the similarity measure used was an euclidian distance between spots. The implementation of DBSCAN and visualisation of final clustered data was done using YALE (Yet Another Learner).

4.3. Results

After comparing the results from two sets of experiments

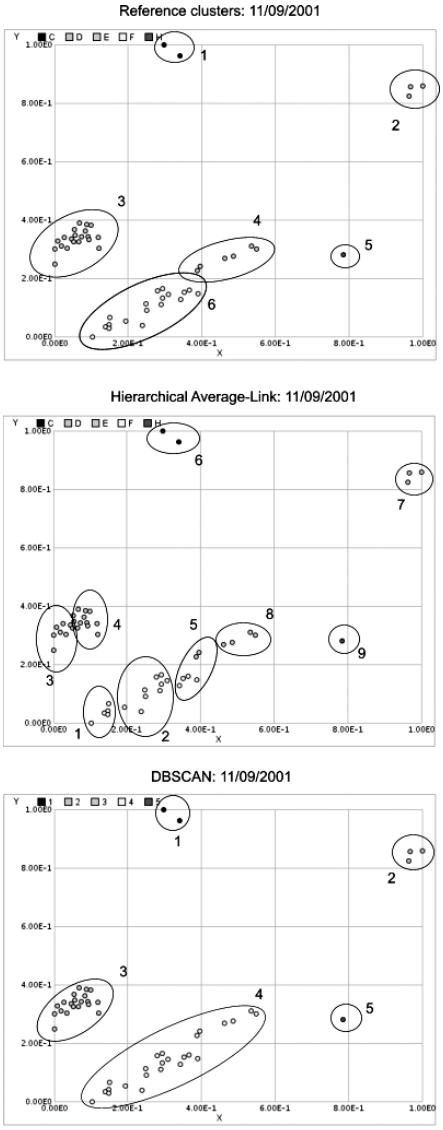

Figure 4. Comparison of clustering results: Hierarchical vs. DBSCAN

it was concluded that the DBSCAN algorithm performed only slightly better compared to the hierarchical. Where it clearly excels was in correctly detecting large, long and elongated sunspot groups. As expected in many cases where the hierarchical algorithm would brake up large sunspot groups into smaller cluster the DBSCAN kept them in one cluster. Also working with DBSCAN input parameters was easier and more predictable than hierarchical's termination condition. The following two examples of sunspot clusters from 11/09/2001 and 23/09/2001 will illustrate the strengths and weaknesses of both methods.

In Figure 4 there are 6 "natural" sunspot groups (labeled 1 to 6 on the reference cluster diagram) with the following classes: C(1) D(2) E(3 and 4) F(6) and H(5). The most interesting of all groups are groups 4 and 6 as they are: 1) very closely located 2) appear to form together a very long and elongated cluster. The hierarchical algorithm, as expected, has broken those two large groups into smaller clusters labeled 1, 2, 5, and 8. Surprisingly it also did not correctly clustered the original group 3 by splitting it into two clusters labeled 3 and 4. In contrast the DBSCAN algorithm had no problem in correctly identifying the original group 3. However it did join the original group 4 and 6 into a single large cluster labeled 4.

Similarly in Figure 5 there are 10 "natural" sunspot groups (labeled 1 to 10). Groups that will likely to cause issues are groups 3 and 5, 9 and 10. The hierarchical algorithm performed better than in the previous example but still failed to detect the original groups 3 and 5 by joining them to form a single cluster labeled 9. Also the original groups 9 and 10 were split into three clusters labeled 7, 8 and 2. The DBSCAN performed more consistently by making similar errors with large clusters as in the previous example. It joined the original groups 9 and 10 into a single large cluster labeled 8. In the same way it joined the original groups 3 and 5 into a single cluster labeled 3. By incrementally lowering the input parameter Eps (to Eps = 0.06) in DBSCAN it was possible to obtain perfect copies of the original groups 3 and 5, 9 and 10 while retaining the remaining correct clusters intact.

In summary the results show that neither single method can be preferred over the other. However the DBSCAN has an edge in dealing with large elongated sunspot groups. Also the algorithm was found to be more predictable, and by varying the input parameters it was possible to obtain perfectly matched clusters. Thus there is some scope for improving on the density based clustering for sunspots data. For example, more attributes could be used to aid clustering such us the areas of spots, number of large spots within a single cluster Large clusters created by DBSCAN, as a result of joining several clusters together, could be subjected to further clustering either using the same algorithm or some type of partitioning algorithm. At this stage the domain

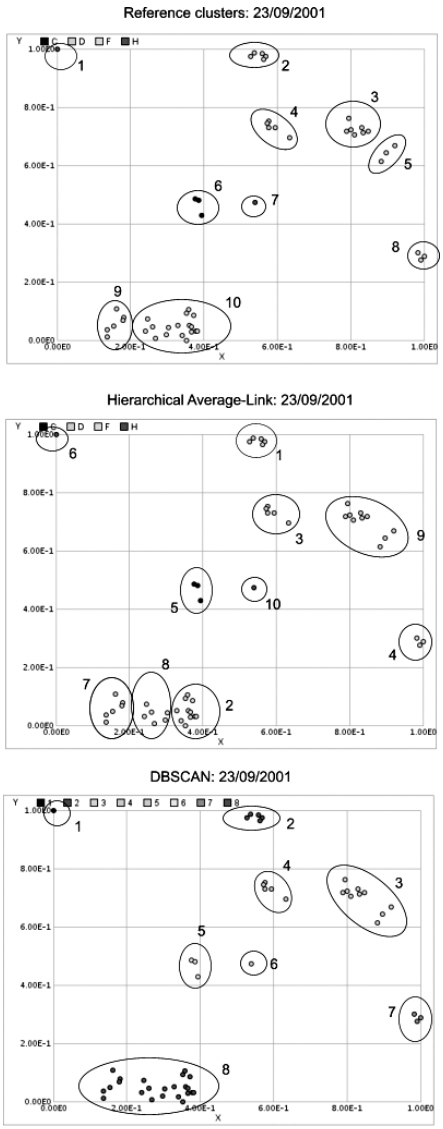

Figure 5. Comparison of clustering results: Hierarchical vs. DBSCAN

knowledge could be used to further refine clusters. Finally distance calculations between spots are currently based on spot centers of gravity. For small spots it is not an issue however for large single spots of elongated shapes the location of spot centers may distort the overall distance calculations.

4.4. Future work

Clearly one single clustering technique is not sufficient to deal with sunspot groups. It may be required to combine several clustering methods in sequence (eg. DBSCAN and k-means for large cluster refinement) or construct a cluster ensemble.

Cluster Ensemble [3] is a powerful method to increase the accuracy of clustering results by combining multiple partitions into a better solution. Constructing a cluster ensemble requires generating multiple clusterings and then combining them into the final clustering. When applying this technique two keys issues needs addressing: 1) How to generate diverse partitions to constitute an ensemble? 2) How to choose a consensus function to combine partitions? Strehl and Ghosh in their recent study presented ways to construct an ensemble of diverse clusterings by utilizing different clustering algorithms and varying subsets of features or patterns. For devising consensus function a variety of method can be employed, including voting, EM algorithm based on mixture model.

Cluster ensemble technique was successfully applied in many fields, including medical diagnostics and gene expression. Experimental results have shown the potential of applying cluster ensemble technique to deal with complex real-world data.

5. Conclusion

The design and workings of the hybrid system for learning sunspot recognition and classification was presented. In past research papers we have shown that it is possible to use the system to successfully recognise and classify sunspots and sunspot groups according to the seven-class Modified Zurich scheme. In this paper a density-based clustering algorithm DBSCAN was compared with an existing hierarchical algorithm implemented in the hybrid system in a task to cluster individual spots (an important step required for sunspot group classification.) The results indicates that although DBSCAN was not substantially better performing it was more consistent and predictable in dealing with large and elongated sunspot groups. Thus a density based method could be further improved.

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

References

[1] Patrick S. McIntosh The Classification of Sunspot Groups Solar Physics, 125 vol. 125 no.2 pp. 251-267, 1990.

[2] Martin Ester, Hans-Peter Kriegel, Jrg Sander, Xiaowei Xu A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996: 226-231.

[3] A. Strehl and J. Ghosh. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3:583-617, 2002.

[4] Anil K. Jain, M. Narasimha Murty, Patrick J. Flynn Data Clustering: Review. ACM Comput. Surv. 31(3): 264-323 (1999)

[5] YALE: Rapid Prototyping for Complex Data Mining Tasks Ingo Mierswa et. al. KDD 2006

[6] Trung Thanh Nguyen, Claire P. Willis, Derek J. Paddon, Sinh Hoa Nguyen, and Hung Son Nguyen. Learning Sunspot Classification *Fundamenta Informaticae, Fundamenta Informaticae, Vol. 72 (1-3) IOS Press, Amsterdam, 2006, pages 295-309.*

[7] Sinh Hoa Nguyen, Trung Thanh Nguyen, Hung Son Nguyen Rough Set Approach to Sunspot Classification Problem In Dominik Slezak, Jingtao Yao, James F. Peters, Wojciech Ziarko, Xiaohua Hu (Eds.): *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31 - September 3, 2005*, Proceedings, Part II. Lecture Notes in Computer Science 3642 Springer 2005. pages 263-272

[8] Trung Thanh Nguyen, Claire P. Willis, Derek J. Paddon, Sinh Hoa Nguyen, and Hung Son Nguyen. Data Mining Approach to Sunspot Classification Problem in *Rough Set Techniques in Knowledge Discovery Workshop at The 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05) in Hanoi, Vietnam 18-20 May 2005.*

[9] 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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] R. J. Bray and R. E. Loughhead. *Sunspots*. Dover Publications, New York, 1964.

[15] 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

[16] Grzymała-Busse J., A New Version of the Rule Induction System LERS Fundamenta Informaticae, Vol. 31(1), 1997, pp. 27–39

[17] The RSES Homepage, `http://logic.mimuw.edu.pl/∼rses`

[18] The WEKA Homepage, `http://www.cs.waikato.ac.nz`

[19] Kjell Inge Malde, CV-Helios Network