On the effectiveness of fuzzy clustering as a data discretization technique for large-scale classification of solar images*

Juan M. Banda, Rafal A. Angryk
Department of Computer Science
Montana State University, Bozeman, MT, 59715 USA

***Abstract-* This paper presents experimental results on the utilization of fuzzy clustering as a discretization technique for purpose of solar images recognition. By extracting texture features from our solar images, and consequently applying fuzzy clustering techniques on these features, we were able to determine what clustering algorithm and what algorithm's initialization parameters produced the best data discretization. Based on these results we discretized some of our texture features and ran them on two different classifiers comparing how well the classifiers performed on our original data versus the discretized data. Our experimental results demonstrate that discretization of our data via fuzzy clustering carries significant potential since on our classifiers produced similar results on the original and the discretized data, and the reduction of storage space achieved through cluster-based discretization has been very significant.**

Index Terms – **discretization, fuzzy clustering, classification, image recognition.**

I. INTRODUCTION

This paper presents our findings on the applicability of fuzzy clustering techniques, to the task of numerosity reduction implemented for the numeric data, characterizing the texture of images of the Sun. The investigation has been conducted in preparation for the upcoming NASA's solar mission named Solar Dynamic Observatory (SDO) [1].

Our overall goal is to build an image-based information retrieval system for astrophysicists, which is expected to work in similar manner as Google does, but with major differences in the character of users' queries and the format of processed data files. We want SDO's researchers to be able to use a query in the form of the image of solar phenomenon to search through the original SDO repository. We want our search engine to be able to quickly point the users toward the solar images that are similar to the initially provided image, ranking the SDO's images based on their relevance to the query.

The task may sound simple, but actual implementation of such a system is far more complex than information retrieval from the text repository. The very first obstacle we encountered is the volume of the NASA's data and its raw character. The SDO mission is scheduled to start taking pictures of the Sun in June 2009, with its Atmospheric Imaging Assembly (AIA) expected to generate one 4096 pixels x 4096 pixels (ten times better resolution than high-definition TV) image every 10 seconds. This will lead to a data transmission rate of $6.76*10^7$ bits per second (i.e. approx. 700 Gigabytes/day) only from the AIA component (the entire mission is expected to be sending about 1.5 Terabytes of data per day, for a minimum of 5 years). Due to the number (and the size) of images that are to be received, and limited computational resources, we are simply not able to constantly work on the raw images. We have chosen to use a number of texture characteristics to represent our images instead. The texture parameters are to be extracted from the images only once, and the above-mentioned image-based retrieval system is expected to be using them instead. The solar images provided by the AIA will be subdivided in 1024 128x128 pixel sections, and for each of the cells multiple texture features will be extracted. This allows for a significant compression of the original data – for instance, if we assume that we have 7 texture features for each of the cells, and that each feature takes 64 bits of memory, with 1024 cells per image, the total space per image will be 56 kilobytes. Since we expect to receive over 8000 images from the AIA module per day, this still generates the growth of our repository at the pace of about 0.5 Gigabyte per day. To compress our data further we decided to consider taking advantage of the discretization of our texture features. If we manage to generate accurate clusters for each of the parameters, we could replace each parameter value with the index of the closest cluster's centroid, which can be then stored in a very small space – leading to further data compression. For instance, if we would have 8 clusters for each of the texture features, we could reduce preliminary storage from 56 kilobytes per image to 1024 cells times 3 bits per feature. This equals about 2.4 kilobytes per image (and approx. 20 megabytes per day).

In this work we investigate the applicability of fuzzy clustering techniques for large scale discretization. We decided to investigate fuzzy techniques, due to their accuracy in domains that do not contain clearly separated clusters. Since the original SDO data is not yet available, we used images of the Sun from the Transition Region and Coronal Explorer (TRACE) [2] mission as a benchmark. Our image repository consisted of 1623 images, 232 images are hand labeled by experts, producing a total of 433 labels in those images. The

* This work was supported in part by the NASA Grant Award No. 08-SDOSC08-0008, funded from NNH08ZDA001N-SDOSC: Solar Dynamics Observatory Science Center solicitation.

types of solar phenomena that we are looking for are: coronal loops, filaments, flares, sun spots and empty-sun (i.e no phenomena exists). All images from the TRACE satellite are in the FITS file format. Our image repository consists of 1623 images.

The rest of the paper is organized as follows: section II constitutes a literature review that presents the background ideas behind our work. An overall description of the image pre-processing, clustering and classification stages is described in section III. Some illustrative experimental results and comments about them are presented in section IV. Conclusions are presented in section V.

II. BACKGROUND

In recent years, multiple attempts were made to automatically identify specific types of phenomena from solar images. Zharkova et al. [3] using Neural Networks, Bayesian interference, and shape correlation, have detected phenomena including sunspots, flares and, coronal mass ejections. Zharkova and Schetinn [4] have trained a neural network to identify filaments within solar images. Wavelet analysis was used by Delouille [5] along with the CLARA clustering algorithm to segment mosaics of the sun. Irbah et al. [6] have also used wavelet transforms to remove image defects (parasite spots) and noise without reducing image resolution for feature extraction. Bojar and Nieniewski, [7] modeled the spectrum of the discrete Fourier transform of solar images and discussed various quality measures. However, we are not aware of any single technique reported as effective in finding a variety of phenomena and no experiments have been performed on the size of repository even comparable to the dataset we are expecting to deal with.

The values produced by our texture feature extraction formulas are not influenced by different orientations of the same kinds of phenomena in different images [8]. Values for these features can also be extracted from the images quickly [9].

Fuzzy clustering has been shown to be effective in clustering medical images [10], and multispectral satellite images [11]. Bezdek [12] developed a fuzzy c-means (FCM) clustering algorithm to allow one piece of data to belong to two or more clusters. Gustafson and Kessel [13] developed a different fuzzy clustering algorithm commonly called GK where each cluster is characterized by its center and a covariance matrix, with the advantage that it can detect clusters of different shapes, but computational costs are more expensive than for FCM. We will compare clustering results provided by the two previously mentioned fuzzy algorithms against two other crisp clustering algorithms: K-Medoids [14] and K-Means [15].

The Fuzzy Clustering and Data Analysis Toolbox [16] was used to apply the previously mentioned algorithms to our image feature vector.

The measures we will use to determine the quality of our clustering have been broadly advertised in literature and have the following characteristics:

Xie and Beni's Index (XB)

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^m \left\| x_j - v_i \right\|^2}{N \min_{ij} \left\| x_j - v_i \right\|^2} \qquad (1)$$

Where c denotes the number of clusters selected, N is the number of elements in our set, v_i is the i-th cluster center.

The index aims to quantify the ratio of the total variation within clusters and the separation of clusters [17]. The optimal number of clusters should minimize the value of the index. This index is bounded between 0 and infinity.

Classification Entropy (CE)

$$CE(c) = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} \log(\mu_{ij}) \qquad (2)$$

Measures the fuzziness of the cluster partition only, which is similar to the Partition Coefficient [12]. CE is bounded between 0 and 1. Where 1 denotes perfect clustering.

Alternative Dunn Index (ADI)

$$ADI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x \in Ci, y \in Cj} \left| \left\| y, -v_j \right\|^2 - \left\| x - v_j \right\|^2 \right|}{\max_{k \in c} \left\{ \max_{x, y \in C} \left\| x - y \right\|^2 \right\}} \right\} \right\} \qquad (3)$$

This is a modified version of the original Dunn’s index [18] that allows its calculation to become simpler and hence less computationally expensive. This index has been originally proposed to be used in the task of identification of "compact and well separated clusters" [7]. The clustering result has to be recalculated as it was a crisp clustering algorithm. A higher value indicates higher quality partitioning.

The main drawback of Dunn's index is computational cost, as calculating it becomes very expensive as *c* and *N* increase, this is the reason the Alternative Dunn Index is usually calculated [16].

There exists only a limited number of datasets that can be used as benchmark for our purposes. No solar images are offered as standard benchmark yet. Since, we were interested in comparing the quality of phenomena recognition using non-discretized texture parameters against a discretized parameter, we decided to try comparative analysis.

Since we have solar images labeled by human experts with the types of phenomena they contain, we decided to evaluate effectiveness of our discretization by comparison of the quality of phenomena recognition when our classifiers are trained on discretized data vs. un-modified texture parameters.

Naïve Bayess and C4.5 classification algorithms will be used for the classification of our datasets [19]. These two algorithms are found in the WEKA [18] data mining software package. To evaluate quality of our classifiers we used the following measures:

$$\text{Recall} = \frac{\left| \{\text{HumanLabel}\} \cap \{\text{ClassifierLabel}\} \right|}{\left| \{\text{ClassifierLabel}\} \right|} \qquad (4)$$

is the probability that a given item from that class will be selected, this measure is bounded by one.

$$\text{F - Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

Is the harmonic mean between precision and recall, this measure is bounded by 1. Precision is defined as the percentage of retrieved items that are relevant.

These measures have been selected to evaluate our results based on the purpose of our investigation to be able to match the labels to our features.

III. DESCRIPTION OF OUR APPROACH

A. Image Pre-Processing

We used grid based image segmentation with 64 cells per image (i.e. 8 x 8 grid), since the image segmentation technique has been proven to produce good results [9]. We calculated 7 features per cell, this give us a total of 103,872 x 7 values to be extracted. The features we extracted are presented in table 1:

TABLE 1
EXTRACTED TEXTURE PARAMETERS

Name	Equation
Mean	$m = \frac{1}{L}\sum_{i=0}^{L-1} z_i$ (6)
Standard Deviation	$\sigma = \sqrt{\frac{1}{L}\sum_{i=0}^{L-1}(z_i - m)^2}$ (7)
Third Moment	$\mu_3 = \sum_{i=0}^{L-1}(z_i - m)^3\, p(z_i)$ (8)
Fourth Moment	$\mu_4 = \sum_{i=0}^{L-1}(z_i - m)^4\, p(z_i)$ (9)
Relative Smoothness	$R = 1 - \frac{1}{1 + \sigma^2(z)}$ (10)
Entropy	$E = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$ (11)
Uniformity	$U = \sum_{i=0}^{L-1} p(z_i) p^2(z_i)$ (12)

Where L stands for the number of elements in the image represented as a vector and z represents a particular element in this matrix.

B. Data Clustering and Discretization

We started our evaluation process with comparing crips vs. fuzzy clustering algorithms. The crisp clustering algorithms applied to our data were K-Means [15] and K-medoids [14]. Fuzzy C-Means (FCM) [12] and GK [13] were our fuzzy clustering algorithms.

For each algorithm the data was normalized to values between 0 and 1 and each feature was clustered individually. We ran each algorithm using 2,4,6,8,10 and 12 as our *cluster numbers* parameter. For preliminary investigation we limited the maximum number of clusters to 12 to be able to store cluster index in 4 bits per each discretized parameter.

We used the selected cluster quality measures (see formulas (1)-(3)) to select the optimal number of clusters, by analyzing which experiments produce the closest to optimal values for each selected measure.

After the clustering was completed we selected the cluster centroids of the best results for our data discretization part of the experiment. The data discretization was performed by replacing original values of texture parameters with the indexes of the closest cluster centroids.

C. Classification of solar phenomena

To perform comparative evaluation of pattern recognition on original vs. discretized/clustered data we selected two commonly used classifiers: (a) Naïve Bayes [20], as the training process is not too expensive computationally, (b) J48 tree better known as C45 [21], which has been used in a similar context for solar image feature classification [9].

We use the F-measure (5) and Recall (4) measure [22] to compare the quality of our phenomena recognition on discretized and non-discretized (i.e. original) texture features.

Using Weka with 10 fold cross validation, the data sets were split into ten equal sized partitions with all classes having approximately the same distribution as the entire set. For each of the classifiers the evaluation was run 10 times for discretized and non-discretized data files. Each time, a different partition was used to evaluate accuracy of classifiers and the remaining nine partitions of the data were used for training [23].

We decided to consider our "discretization via clustering" to be effective when accuracy of the classifiers did not decreased for discretized data (when compared against original values of texture features)

IV. EXPERIMENTAL RESULTS AND COMMENTS

A. Clustering

After completing our experiments we were able to determine which feature under what clustering algorithm, and at what number of clusters produced the best results for our selected cluster quality measures.

Due to the limited size of this paper and complexity of our evaluation (evaluation of quality of clustering needed to be followed by evaluation of image recognition to asses effectiveness of our clustering-based discretization approach), we decided to limit this part of our presentation to two of our seven features (listed in table 1): *third moment* and *fourth moment*. We have chosen these 2 features because they generated the most interesting results. We present figures 1,2,3 and 4 showing the clustering results of hard clustering algorithms versus fuzzy clustering algorithms for our selected quality measures values for each number of clusters. Based on these results we decided to apply the FCM algorithm on several features of our feature vector, but using different parameters for the algorithm, and also changing the algorithm's shape of the default membership function to a Gaussian shaped membership function, since this has been believed to produce better result than traditional triangular membership functions

Based on these results, we can conclude that the *third moment* feature produced the best results using the Fuzzy C-Means algorithm with 6 clusters. This is based on the XB and ADI measure values as we can see in figures 1 and 2.

Fuzzy C-Means for the *fourth moment* feature produced the best results on the ADI and CE measure scales, as shown in figures 3 and 4. This time the clearest grouping of data values was achieved for 4 clusters.

In all investigated cases fuzzy clustering algorithms show better (or at least equally good clustering, when compared against crisp methods. We believe that fuzzy clustering is a very good fit for the processing of solar images, where regions of interest have more gradual (fuzzy) borders than in the case of home-made pictures, which often contain sharp objects (e.g. contours of house with blue sky background) more often.

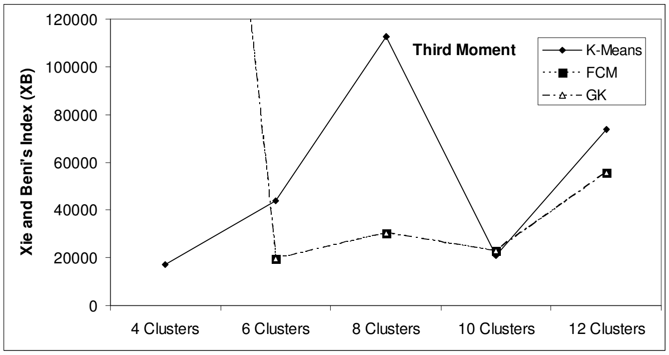

Figure 1. XB measure values for the *third moment* feature. Smaller is better. K-medoids did not produce values for this measure.

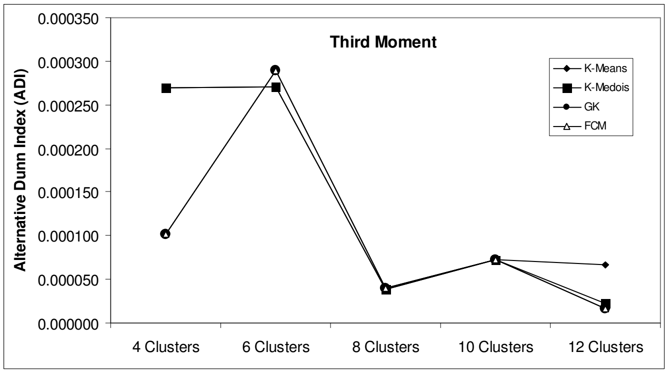

Figure 2. ADI measure values for the *third moment* feature. Bigger is better.

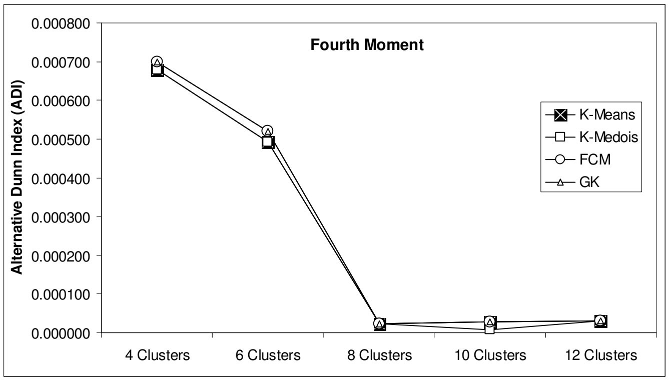

Figure 3. ADI measure values for fourth moment feature. Bigger is better.

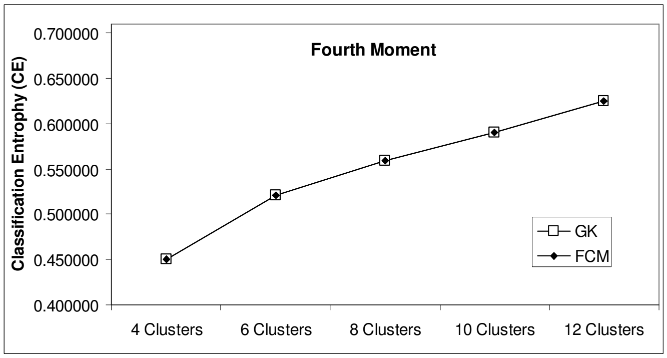

Figure 4. CE measure values for the *fourth moment* feature. Smaller is better. Since CE is a fuzzy clustering measure, it did not yield any results for crisp clustering algorithms.

To further improve the quality of our clustering, we decided to investigate the behavior of our best clustering algorithm (i.e. Fuzzy C-Means) with different fuzzification parameters. We used 4 settings of fuzzification parameters as shown in table 2.

TABLE 2
FUZZY C-MEANS CLUSTERING PARAMETERS

Name	Parameters
FCM 1	m = 1.5 (more crisp)
FCM 2	m = 2 (default)
FCM 3	m = 3 (more fuzzy)
FCM 4	Gaussian with m = 3

The parameter m indicates the weighting exponent which determines the fuzziness of the clusters. As m approaches one from above, the partition becomes hard, the further it goes to infinity the partition becomes more fuzzy [16].

We also modified the Fuzzy C-Means algorithm standard membership function to investigate the applicability of a Gaussian membership function. Although the Gaussian curve is more computationally expensive to process than triangular functions, we wanted to investigate their influence on the quality of solar data organization.

We ran our experiments with the table 2 settings for 2,4,6,8,10 and 12 clusters. Figures 5,6,7 and 8 represent our results.

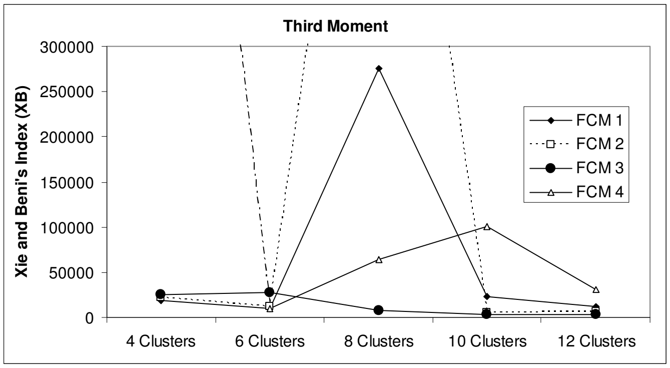

Figure 5. XB measure values for the *third moment* feature. Smaller is better

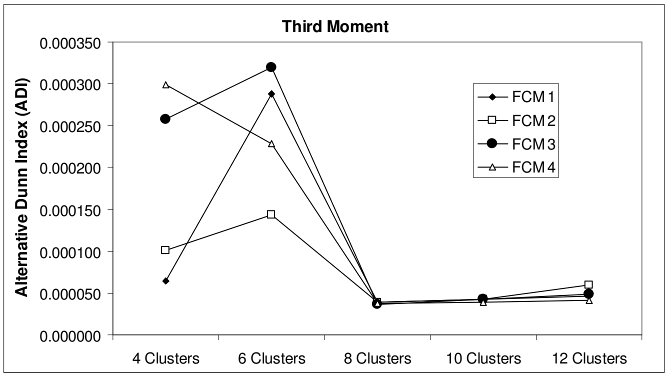

Figure 6. ADI measure values for third moment feature. Bigger is better.

Figures 5 and 6 show that some of our parameter modifications provide small improvements over the default settings (FCM2) in the values of our quality measures. This makes them more useful for the discretization of our data. Once again we have confirmation that there are 6 clusters in the *third moment* feature. Gaussian shaped membership function did not generate high enough improvement of clustering quality to justify their use on larger SDO image corpora. As we can see again, flattening of membership function allowed for better clustering than using sharp borders.

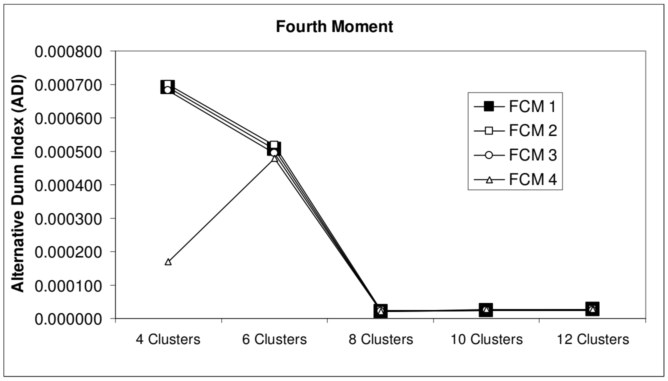

Figure 6. ADI measure values for fourth moment feature. Bigger is better.

Figure 6 shows that the majority of our parameter modifications produce very small improvements for the ADI index, when compared against the default settings (FCM2). In almost all of the cases any results generated by the Gaussian membership functions occurred at the bottom of the pack.

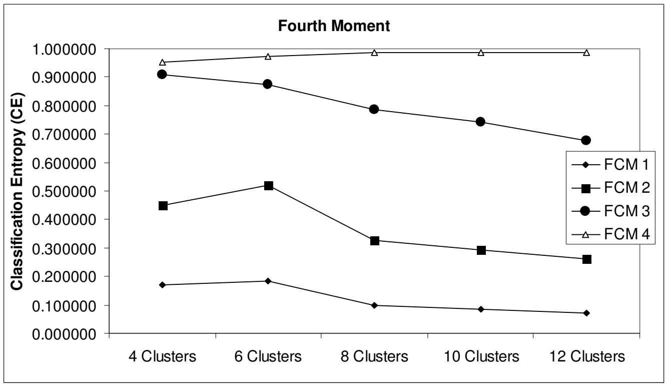

Figure 7. CE measure values for fourth moment feature. Smaller is better.

In figure 7 is the only place where we see considerable improvement of our clustering results for our parameter changes. Once again the results indicate that Gaussian shaped membership functions are not a good fit for our solar data.

While K-medoids provides close results to FCM, we were looking to determine what clustering algorithm provided the best overall results for the selected number of clusters. On top of FCM performing slightly better, this algorithm allowed us to modify its membership function in an attempt to generate better results.

B. Classification of solar phenomena and comparative evaluation

After determining the clusters among our data, we were able to move toward comparative evaluation of our clustering-based data discretization process. The approach we took is quite straightforward. First we decided to check how accurate recognition of solar phenomena we can get using non-discretized values of our texture features, and then we could compare these results with accuracy of (the same) classifiers but trained (and tested) on discretized data. To independently evaluate quality of discretization for each of the features, we discretized (i.e. replace original values of the investigated feature, with its clusters centroids) only one feature at the time.

The datasets we used for comparative evaluation are referenced the following way:

- **Original** is our original labeled image data. Seven features where extracted but no discretization was performed.
- **FCM 1 M3** is where the FCM 1 setting where used to discretize the data for the *third moment* feature
- **FCM 2 M4** is where the FCM 2 setting where used to discretize the data for the *fourth moment* feature

We ran our two previously-mentioned classifiers (Naïve Bayes and J48) on the three datasets we described in this section, for the purpose of comparative evaluation.

TABLE 3
ACCURACY OF SOLAR PHENOMENA RECOGNITION ON DISCRETIZED VS. ORIGINAL DATA

	NaiveBayes	C45
Original F-Measure	0.7510	0.7890
FCM 1 M3 F-Measure	0.7537	0.7886
FCM 2 M4 F-Measure	0.7540	0.7891
Original Recall	0.8700	0.9765
FCM 1 M3 Recall	0.8783	0.9759
FCM 2 M4 Recall	0.8783	0.9762

Table 3 shows the F-Measure and Recall measures. Based on some of these results, we are able to discriminate which classifiers produce the best classification results and which can be discarded and also estimate how discretization of individual features is influencing the values of the quality of solar phenomena recognition.

After analyzing our classification results we can say that C4.5 classifier produced better results than Naïve Bayes. Also we could say that the classifiers on our discretized features produced almost identical values than our original labeled data, which leads to the conclusions that we can use the clustering-based discretization of the data and obtain almost the same quality of solar phenomena recognition.

V. CONCLUSIONS

During our clustering experiments we encountered that for two of our features presented the Fuzzy C-Means algorithm provided the best clustering according to our clustering measures. This motivated us to expand our experiments further, by modifying the fuzzy weight parameter as well as changing our membership functions. These experiments provided interesting results showing that they could improve our clustering by a small factor, so we decided to use these results for discretization of texture features during our classification experiments.

Our results are, in general, marginally better for the discretized data because after discretizing by cluster centers we gathered all data points that were at the borders of our clusters to the cluster center. The clustered data has less variance resulting in a more linear model of the classifiers, since we now have pure classes with pure borders.

Our classification results showed that with discretized data we obtain similar classification quality, allowing us to be able to discretize our data, and therefore solve our storage problem providing a considerable reduction in the size of the data that will need to be stored the SDO repository.

VI .ACKNOWLEDGEMENT

The authors would like to state their gratitude to all the persons involved in the TRACE mission [2], who provided the labeled TRACE images for preliminary evaluation.

In addition, the authors would like to thank the SDO mission for partially sponsoring this research.

REFERENCES

[1] SDO | Solar Dynamics Observatory [Online], Available: http://sdo.gsfc.nasa.gov/. [Accessed: Feb 12, 2008]

[2] TRACE On-line (TRACE) [Online], Available: http://trace.lmsal.com/. [Accessed: Feb 12, 2008]

[3] V. Zharkova, S. Ipson, A. Benkhalil and S. Zharkov, “Feature recognition in solar images," *Artif. Intell. Rev.*, vol. 23, no. 3, 2005, pp. 209-266

[4] V. Zharkova and V. Schetinin, “Filament recognition in solar images with the neural network technique," *Solar Physics*, vol. V228, no. 1, 2005, pp. 137-148, [Online]. Available: http://dx.doi.org/10.1007/s11207-005-5622-1

[5] V. Delouille, J. Patoul, J. Hochedez, L. Jacques and J.P. Antoine ,“Wavelet spectrum analysis of eit/soho images," *Solar Physics*, vol. V228, no. 1, 2005, pp. 301-321, [Online]. Available: http://dx.doi.org/10.1007/s11207-005-5620-3

[6] A. Irbah, M. Bouzaria, L. Lakhal, R. Moussaoui, J. Borgnino, F. Laclare and C. Delmas, “Feature extraction from solar images using wavelet transform: image cleaning for applications to solar astrolabe experiment.” *Solar Physics,* Volume 185, Number 2, April 1999 , pp. 255-273(19)

[7] K. Bojar and M. Nieniewski. “Modelling the spectrum of the fourier transform of the texture in the solar EIT images”. *MG&V* 15, 3, January 2006, pp. 285-295.

[8] R.M Haralick, K. Shanmugam and I. Dinstein, “Textural Features For Image Classification,” IEEE *Transactions on Systems, Man, and Cybernetics,* Volume: SMC-3, No. 6, Nov. 1973, pp 610- 621.

[9] R. R. Lamb, “An Information Retrieval System For Images From The Trace Satellite,” M.S. thesis, Dept. Comp. Sci., Montana State Univ., Bozeman, MT, 2008.

[10] A. Lorette, X. Descombes and J. Zerubia, "Fully Unsupervised Fuzzy Clustering with Entropy Criterion," *15th International Conference on Pattern Recognition* (ICPR'00) - Volume 3, 2000, pp.3998.

[11] L. Podenok and R. Sadykhov, "Multispectral Satellite Image Segmentation Using Fuzzy Clustering and Nonlinear Filtering Methods," *International Machine Vision and Image Processing Conference*, 2008, pp.43-48

[12] J. C Bezdek: “Pattern Recognition with Fuzzy Objective Function Algorithms”, *Plenum Press*, New York, 1981.

[13] D.E. Gustafson and W.C Kessel, “Fuzzy clustering with a fuzzy covariance matrix,” *Proc. IEEE Conf. Decision Contr.*, San Diego, CA, 1979

[14] L. Kaufman and P. J. Rousseeuw, “Clustering by means of medoids.” Statistical Data Analysis based on the L1 Norm, *Elsevier*, 1987, pp. 405-416.

[15] J.B MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, CA, 1967, pp. 281-297

[16] B. Balasko, J. Abnyi and B. Feil, *Fuzzy Clustering and Data Analysis Toolbox for Use with Matlab*, University of Veszprem, Veszprem Hungary, 2005, [Online] Available: http://www.fmt.vein.hu/softcomp/fclusttoolbox/

[17] X. L. Xie and G. A. Beni. Validity measure for fuzzy clustering. *IEEE Trans. PAMI*, 3(8):841-846, 1991.

[18] P. Reutemann, B. Pfahringer and E. Frank. (2004) *Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners.* 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag."

[19] J.M. Keller and S. Chen. “Texture description and segmentation through fractal geometry.” *Computer Vision, Graphics, and Image Processing,* 45, 1989, pp. 150-166.

[20] P. Domingos , M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning* 29, 1997, pp 103–130.

[21] J. R. Quinlan. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research 4, 1996, pp 77-90.

[22] C. J. van Rijsbergen. Information Retireval. Butterworths, London, 1979.

[23] I. H. Witten and E. Frank, Data Mining:: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kaufmann, 2005.