Hailstone classifier based on Rough Set Theory

To cite this article: Huisong Wan et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 231 012087

View the article online for updates and enhancements.

You may also like

- Clustering Reduction Method Analysis of Rough Set and Decision Tree based on Weight Matrix Analysis
  Lan Wang and Yi Zhang
- Supplier Selection With Gray Based Rough Set Theory Method (A Case Study: Pharmaceutical Installation Of RSU Grand Medica Tanjung Anom, Medan)
  A Bakhtiar, Y S T Siahaan and A Susanty
- Research on Intelligent Scale Pedestrian Navigation Map Based on Rough Set Rules
  Chaode Yan, Xiao Liu, Jinxi Guo et al.
Hailstone classifier based on Rough Set Theory

Huisong Wan, Shuming Jiang, Zhiqiang Wei, Jian Li and Fengjiao LI
The information Research Institute of Shandong Academy of Sciences, Jinan, 250014, China

Abstract. The Rough Set Theory was used for the construction of the hailstone classifier. Firstly, the database of the radar image feature was constructed. It included transforming the base data reflected by the Doppler radar into the bitmap format which can be seen. Then through the image processing, the color, texture, shape and other dimensional features should be extracted and saved as the characteristic database to provide data support for the follow-up work. Secondly, Through the Rough Set Theory, a machine for hailstone classifications can be built to achieve the hailstone samples’ auto-classification.

1. Introduction
Hailstone is the main disaster in the Meteorological area. Every year, hailstone brings huge losses to agriculture, architecture, communications, electricity, transportation and people's lives and property. In recent years our country paid high attention and gave strong supports to the forecast and monitoring of the hailstone.

Rough Set theory was proposed by Professor Z.Pawlak, a scientist in Poland in 1982. This theory is mainly used to deal with the vagueness and uncertainty of the mathematical problem, which is mainly used to find the rules from a large number of the incomplete data sets. Therefore, Rough set theory is the basis of the knowledge acquisition, machine learning and reasoning of uncertain information.

The study of the hailstone forecast only pays attention to whether the hailstone fall or not, but do not pay close attention to the style of the falling hailstone. At the same time, the style of hailstone is an important point for the prevention of the hailstone. Therefore, the study of this paper is very meaningful.

2. The conversion of the radar image
The feature extraction of the image is the key point of the construction of the characteristic database, which is also the foundation of the construction of the hailstone classifier. The most important task of the feature extraction of the image is to complete the conversion of the radar image.

The weather radar system is the important tool of the monitoring and early warning of the strong convective weather (hailstone, tornadoes and flash floods). The new generation of the weather radar is made up of a certain number of subsystems. The subsystems are: Radar Data Acquisition (RDA), Radar Product Generation (RPG), Principal User Processor (PUP) and the communication lines between them.

The data record of RDA mainly includes two parts: The first level data: The original analog data which were received by the receiver and would not be saved. The second level data (base data): The base data are the highest temporal accuracy data (the reflectance factor, mean radial velocity and the velocity spectrum width) which are generated by the signal process.
The data base is the original data. Since the base data cannot be displayed and follow-up treated, so it is necessary to transform data base to BMP format.

In order to better deal with the base data, the class CBASEDATE was structured, which was used for:

1. Extracting the information of the base data and saving.
2. Getting the information which we are interested through calculating the information of the database.
3. Translating the base data into 8-bit bitmap with the width of 512 and the height of 512. In the radar image, the difference of reflectance (intensity) is reflected by different colors, so we need to define the color table.

3. The construction of characteristic database

Hailstone which is a typical convective weather has its own image features. A complete hailstone process has three stages: the formation stage, the mature stage, the landing stage and the extinction phase.

The radar image feature is the underlying data source of follow-up work, the quality of the selection and extraction of the radar feature is directly related to the quality of the hailstone classifier.

3.1 The color feature

The color feature describes the surface property of the scenery, so the color feature is the most widely used feature. Color feature is a global feature which is not sensitivity to the changes in size; therefore, the color feature has a high robustness. Because the color feature is correspond with the radar reflectivity directly, and the number of color is only fifteen, so the color feature is easy to extract in this paper.

The common color feature representation – color histogram is used to describe the color feature. Firstly, the color histogram was got by counting the proportion of each color in the image, and then extracted the five features: mean, variance, skew, kurtosis, the proportion of echo.

According to the knowledge of the weather forecast, the proportion of the high echo in the image is the important parameter in the hailstone forecast, therefore, another two features were extracted, which were the proportion of 45 and 55 higher color frequency in the total frequency.

3.2 Texture feature

Texture feature is another common global feature, which is also the low-level feature and only response the surface characteristics. The texture described the relative distribution of every gray-scale image pixel, and it also reflected the changes of gray-scale or 2D colors. Therefore, the texture of image could be measured by texture features. At present, satellite cloud textures have been analyzed successfully to realize the identification of various clouds (such as cumulonimbus, stratus, and cirrus) and strong convective weathers like typhoon automatically. In this paper, the method of gray level co-occurrence matrix (GLCM) in statistical was introduced to analyze radar image from statistical texture features. We extracted seven more useful texture feature—TexEgy, TexInt, TexCl, TexEpy, TexLc, TexAl and Gradsufy.

3.3 The shape feature

The shape feature is one of the important features of describing image. It is related with the object or the target, and has a certain semantic meaning. Therefore, the shape feature can be regarded as the higher layer of image characteristic description than color feature and texture feature.

This paper adopted several common geometrical parameters to describe the shape of the echo in the radar image: area, length, circular degree, equivalent standard nuclear area, the area of warning region, the length of the hook, the shape index of the hook.

In this paper, Threshold segmentation and Region labeling method were adopted on the bitmap, then the region of interest in the image were got, and their color, texture and shape feature were ex-
tracted which were stored in the database. In this paper, the database contains the information of 576 regions.

4. The design of hailstone classifier based on Rough Set theory

In this paper, rough set theory and method is applied to the mining of hail rules. The block diagram is shown in Fig. 1.

![Diagram](image)

Figure 1. The block diagram of the rough set theory

4.1 The decision table

In the decision table, each column represents a characteristic attribute of the object. The characteristic attributes are divided into conditional attributes and decision attributes. The condition attributes of this article is the 17 dimensional feature extraction of hail (lots of experiments show that using all 24 dimensional features cannot achieve good results, so we removed 7 features), the decision attribute for hail state, this paper is divided into three states: 1, 2, 3 respectively, the formation of hail during the landing phase, and death phase. Each row of the decision table represents a specific sample. From the previous analysis, we know that each attribute is an equivalence relation in the decision Table 1.

| Number | Conditional attributes | Decision attributes |
|--------|------------------------|--------------------|
|       | TexAverageLum | TexEnergy | Cmean | Ckurt | Cskew | ...... |       |
| 1      | 2.33945 | 0.06871 | 23.08550 | 715318 | 6205.37 | ...... | 1 |
| 2      | 3.20444 | 0.07550 | 24.16670 | 883302 | -312.43 | ...... | 1 |
| 3      | 1.92568 | 0.14195 | 21.42620 | 606658 | 10233.17 | ...... | 1 |
| 4      | 2.26832 | 0.07530 | 26.32420 | 1622540 | 14997.14 | ...... | 2 |
| 5      | 2.08232 | 0.08356 | 29.18880 | 2953780 | 14217.01 | ...... | 2 |
| 6      | 1.95954 | 0.10732 | 25.24680 | 1133600 | -9408.02 | ...... | 2 |
| 7      | 2.25470 | 0.06997 | 30.03500 | 2953570 | 16292.01 | ...... | 3 |
| 8      | 1.62149 | 0.08767 | 29.08330 | 2204550 | -2213.15 | ...... | 3 |
| 9      | 1.65724 | 0.09634 | 25.06000 | 1133420 | -9765.12 | ...... | 3 |
| 10     | 1.96474 | 0.08162 | 21.27100 | 383765 | 21058.37 | ...... | 3 |

4.2 The completion of the data

Due to various reasons, the radar data we get is not complete, and there are some default attribute values in tables, if you remove these default values, it will result in a waste of resources, and will lose some information hidden in the default data, so it is very necessary to make up the data. The commonly employed approaches include Mean Completer, Combinatorial Completer and ROUSTIDA. In this paper, the method of Mean Completer is used to complete the missing data.

4.3 The discretization of the data

Generally speaking, rough set theory cannot deal with the continuous attributes directly. The discretization is to set a number of discrete points in the range of the continuous attributes, and the continuous attributes are divided into a number of discrete intervals, and the continuous values of each sub interval are represented with different symbols or branches. The commonly used methods of discretization are Boolean reasoning algorithm, equal frequency discretization and so on. In this paper the equal frequency discretization is used. The basic idea of equal frequency discretization: each feature is arranged in the ascending order, and the test sample is divided into k sub intervals according to the frequency set K, and each sub interval contains the same number of samples. After many experiments,
this article will be divided into 8 parts of equal frequency. The discrete points of some condition attributes are shown in Table 2.

Table 2. The discrete points of some condition attributes

| Conditional attributes | The discrete point 1 | The discrete point 2 | The discrete point 3 | The discrete point 4 | The discrete point 5 | The discrete point 6 | The discrete point 7 |
|------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| TexAverageLum          | 1.96257              | 2.73406              | 3.30296              | 3.87626              | 4.38358              | 5.10747              | 5.85470              |
| TexEnergy              | 0.0406               | 0.04839              | 0.0576               | 0.0699               | 0.087                | 0.11                | 0.156                |
| Cmean                  | 26.122               | 29.387               | 31.468               | 33.933               | 36.347               | 38.74                | 42.166               |
| Ckurt                  | 725875               | 1403200              | 2028990              | 3091490              | 4603380              | 6614660              | 11217000             |
| Cskew                  | -10059.8             | -5536                | -4884.3              | -335.35              | 4962                 | 5651.1               | 10097.5              |

4.4 The reduction of the data
The commonly used attribute reduction algorithms are the genetic algorithm and Johnson algorithm. The genetic algorithm is to simulate the competitive mechanism of "survival of the fittest", and to search and find the optimal path through the iterative method; Johnson algorithm is good at discovering the minimum length of the implicit rules, and the speed is very fast.

In this paper, Johnson algorithm is used to reduce the data. This algorithm is mainly realized by the greedy algorithm. In this paper, Johnson algorithm is used to reduce the 489 samples, and the results of partial reduction are shown in Table 3.

Table 3. The results of partial reduction

| The results of reduction | The results of reduction | The results of reduction | The results of reduction |
|-------------------------|-------------------------|-------------------------|-------------------------|
| {CHpercent}             | {Cmean, CSfCompactness} | {Cvariant, CArea}       | {TexAverageLum, TexInertia, CCorepercent} |
| {TexCorrelation}        | {Cskew, CSfCompactness} | {Cmean, Cvariant}       |                         |
| {TexAverageLum}         | {TexEntropy, Cskew}     | {CHpercent, CArea}       |                         |
|                         |                         |                         |                         |

4.5 The generation of the rule
According to the result of reduction, 126 rules are generated, and some results are shown in Table 4.

Table 4. The part of the rule

| Rule                                                               | LHS Support | RHS Support | RHS Accuracy | LHS Coverage | RHS Coverage |
|--------------------------------------------------------------------|-------------|-------------|--------------|--------------|--------------|
| CHpercent(6) => 3 or 2 or 1                                        | 62          | 53,6,3      | 0.855, 0.0968, 0.048 | 0.127        | 0.17, 0.05, 0.048 |
| CHpercent(7) => 3 or 1 or 2                                        | 61          | 49,4,8      | 0.8, 0.07, 0.013   | 0.1247       | 0.157, 0.06, 0.07 |
| TexGradsUniformity(5) and CHpercent(7) => 3                        | 10          | 10          | 1             | 0.02         | 0.032        |
| TexAverageLum(4) and TexInertia(6) => 2                             | 6           | 6           | 1             | 0.012        | 0.0517       |
| TexAverageLum(3) and TexEnergy(4) and Cvariant(7) => 2             | 6           | 6           | 1             | 0.012        | 0.0517       |
| TexAverage-                                                         | 7           | 7           | 1             | 0.014        | 0.11         |
The rule is the objective model obtained by reduction of the data, for example through a single reduction CHpercent, it can be obtained: CHpercent (6) => 3 or 2 or 1. When the discrete value of CHpercent is 6, the coverage is 62, and the hailstone may be three kinds of circumstances, but the reliability of each case is 0.855, 0.0968, 0.048. So when CHpercent=6 occurs, we give priority to 3 (demise), while ignoring the other two possibilities.

In this paper, the conditional attributes involved in the rules are mainly two items, but there are also a small number of cases with three items. The rules are arranged in the form of knowledge base, which is convenient for later use.

In this paper, the 126 rules are used to verify the results of the 86 hailstone samples, and the results are shown in Table 5.

| The actual style | 1   | 2   | 3   | unknown |
|------------------|-----|-----|-----|---------|
| The result of the classifier | 0   |    | 0   | 2       |
|                   | 2   | 1  | 2   | 0       |
|                   | 3   | 2  | 0   | 54      | 0       |

According to the above table, we can see that the number of the missing forecast is 2, and the proportion is 2.3%, and the number of the accurate forecast is 56, and the proportion is 65%. The number of the false alarm forecast is 28, and the proportion of 32%.

5. Conclusions
The database of the radar image feature was constructed. It included transforming the base data reflected by the Doppler radar into the bitmap format which can be seen. Then through the image processing, the color, texture, shape and other dimensional features should be extracted and saved as the characteristic database to provide data support for the follow-up work. The Rough set theory was used for the construction of the hailstone classifier. This paper made a bold attempt to the hailstone classifier. This study of this paper is a challenging and promising work, and is worth in-depth studies in the future.

References
[1] Wan Huisong, Lu Zhiying, Jiang Shuming, Zhang Yuanyuan. Research on the comprehensive Forecast of Hailstone. GCIS 2012
[2] Zhiying Lu, Lei Wang, Hongmin Ma. Hailstone Detection Based on Image Mining. IEEE: 2008
[3] Xiangyang Liu, Huisong Wan, Yuanyuan Zhang. Hailstone Classifier Based on Back Propagation neural Network. ICMCE 2013
[4] J. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski (2000): Rough set algorithms in classification problem. In L. Polkowski, S. Tsumoto, and T. Lin, editors, Rough Set Methods and Applications, Physica-Verlag, Heidelberg New York., 49–88.
[5] Zhiying Lu, Huisong Wan, Huizhen Jia, Knowledge Discovery of Hailstone Echo Time Series Base on Mutual Information Network Model, Fuzzy Systems and Knowledge Discovery, 2011, FSKD’11.
[6] Luo bing, Li huasong, The principle and application of artificial intelligence, Beijing, Mechanical press, 2011 (in Chinese)
[7] Wang wenjie, Ye shiwei, The principle and application of artificial intelligence, Beijing, Post&Telecom Press, 1999 (in Chinese)