CLOUD MODEL AND HIERARCHICAL CLUSTERING BASED SPATIAL DATA MINING METHOD AND APPLICATION

Kun Qin[a,*], Min Xu[a], Yi Du[b], Shuying Yue[a]

[a]School of Remote Sensing Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, 430079, China - qinkukn163@163.com, xumin818@126.com, yueshuying_2005@sina.com
[b]Network Technology Management Center - yidu@sina.com

Commission II, WG II/2

KEY WORDS: Cloud Model, Hierarchical Clustering, Data Mining, Weather Classification

ABSTRACT:

The paper proposed a method of weather classification based on cloud model and hierarchical clustering. Firstly, according to the weather sampling data from FY-2C images, cluster the brightness values of five wave bands by hierarchical clustering approach based on cloud model, and generate the centres and the parameters of clusters of image features in different rainy weather types. Secondly, through the precondition cloud generator, compute the membership degree of each pixel of FY-2C images to each cluster which represented by cloud model. Lastly, determinate the precipitation weather type of corresponding weather stations. The experiments validate the proposed method.

1. INTRODUCTION

With the rapid development of spatial information science, how to discover knowledge from spatial database and promote the intelligent spatial information processing becomes an important research direction (Li and Cheng, 1994; Koperski et al., 1996; Li et al., 2006). Recently, the research of spatial data mining has gradually turned to the application research, such as marine ecological research, remote sensing, climate research and so on (Su et al., 2004; Qin, 2004; Yang et al.,2007). Satellite cloud image is an important kind of remote sensing images. It can reflect both the distribution and the change of large-scale cloud images. Particularly, it can be used to analyze and predict the precipitation according to the reflection of weather cloud images to the temperature of cloud top. Therefore, weather cloud images have played an important role in the fields of meteorology, hydraulics.

The satellite cloud image data are important information for weather analysis and prediction. In recent years, with the development of the digital satellite cloud images, the quantitative analysis on satellite cloud images has become an important research direction to some of the meteorological researchers (Wang et al., 2005). For example, Koffler proposed a method to distinguish clouds from earth surface based on threshold approach (Koffler, 1973). Desbois proposed the concept of spectrum characteristics space and proposed the method of cassette classification which promotes the development of simple threshold method (Desbois, 1982). Welch et al. researched the method of cloud classification according to texture characteristics (Welch et al., 1989).

The research on the relationship between satellite cloud images and precipitation has been carried out for many years. The researchers used various methods to analyze the information of satellite cloud images, expected to find the relationship between cloud images and precipitation, and finally realized the precipitation prediction based on cloud images. Chen and Yu estimated the plum rainfall using cloud images by establishing the regression equation of cloud types and precipitation, and classified the precipitation using bispectrum cloud image threshold and units feature space (Chen and Yu, 1994; Yu,1998). Wang et al. estimated the precipitation on stratus cloud and convective cloud according to the relationship between the core temperature of top convective cloud and precipitation gained by one-dimensional model of cloud on the basis of the geostationary satellite cloud image classification (Wang, 1998). Li et al. finished the experiment which pre-processes the satellite images based on wavelet analysis, combine with the meteorology radar data, and carry out the test of precipitation estimated by the method of neural network (Li et al., 2000). Shi et al. researched on the precipitation classification using the method of image segmentation combine with artificial neural network (Shi et al.,2001).

Clustering is a very important technique of data mining. Using the approach of clustering to research the relationship between satellite cloud images and precipitation has been increasingly attracted the attention of corresponding scholars. Hong et al. developed an idea to combine FCM, GA with FSC mutually, and use them to cluster the high-dimensional features of GMS-5 images. Then, calculate the distance between these samples and clustering centres to determine their classifications. The type of the pixels in original cloud images can be found out which group in high-dimensional feature spaces the pixels belong to. So that it can make sure its weather area to accomplish the automatic classifications of the weather area (Hong et al., 2006).

The paper proposes a method of weather classification based on cloud model and hierarchical clustering. According to weather sampling data from FY-2C image, cluster the brightness values of five wave bands by the method of cloud model based hierarchical clustering, and generate the clusters of image features under the condition of different rainy weather types

[*] Kun Qin, email: qinkun163@163.com

which composed the precipitation weather judgment model of cloud images. Then compute the membership degree of each pixel in the real time cloud images of FY-2C to each cloud model through the forward cloud generator. Finally, determinate the precipitation weather types of corresponding weather stations. the precipitation weather type of corresponding site belongs to can be ensured, and the real-time precipitation weather judgment can be realize.

2. THE PRINCIPLES AND METHODS

2.1 Cloud Model and Its Key Technologies

2.1.1 The Concept of Cloud Model:

Cloud model is a conversion model with uncertainty between a quality concept which is expressed by natural language and its quantity number expression (Li and Du, 2007). If *U* is a quantity domain expressed with accurate numbers, and *C* is a quality concept in *U*, if the quantity value $x \in U$, and *x* is a random realization of the quality concept *C*, $\mu(x)$ is the membership degree of *x* to *C*, $\mu(x) \in [0,1]$, it is the random number which has the steady tendency:

$$\mu : U \to [0,1] , \ \forall x \in U , \ x \to \mu(x)$$

The distribution of *x* in domain is called cloud model, which is briefly called cloud, each *x* is called a cloud drop.

2.1.2 The numerical characteristics of cloud model:

The numerical characteristics of cloud model are expressed with Expectation *Ex*, Entropy *En* and Super-entropy *He*, and they reflect the whole characteristics of the quality conception *C*. Expectation *Ex* of the Cloud drops' distribution in domain, is the point which can best represent the quality concept, reflect the cloud centre of gravity of cloud drops of the concept (Li and Du, 2007). Entropy *En* is the uncertainty measurement of the quality concept, is decided by the random and fuzziness of the concept, it reflects the connection between the fuzziness and the random. Entropy *En* is a random measurement of the quality concept, reflects the discrete degree which can represent the quality concept; in another aspect, it is the measurement of fuzziness, and it reflects the value range which can be accepted by the concept of the cloud drop (Li and Du, 2007). Use Entropy *En* the same numeric characteristic to reflect fuzziness and random, and it embodies the connection between each other. The super-entropy *He* is the uncertain measurement of entropy, namely the entropy of the entropy. It reflects the coagulation of uncertainty of all points which representing the concept in the number domain, namely the coagulation degree of cloud drop. The size of super-entropy reflects the discrete degree and thickness of cloud indirectly. The super-entropy is decided by the random and fuzziness of entropy together.

2.1.3 Normal cloud model:

The normal distribution has the characteristic of universality, the expectation curve of cloud of the quality knowledge about the social and natural sciences all approximately to the normal or the pan-normal distribution. The normal cloud model is the most basic cloud model, and it is an efficient tool to express language atom (Li and Du, 2007). The expectation curves equation of cloud decided by expectation and entropy is as follows.

$$y = \mathrm{e}^{-\frac{(x-E_x)^2}{2E_n^{\ 2}}}$$

Given E_x=0， E_n=3， H_e=0.3， n=10000. It is illustrated in Figure 1.

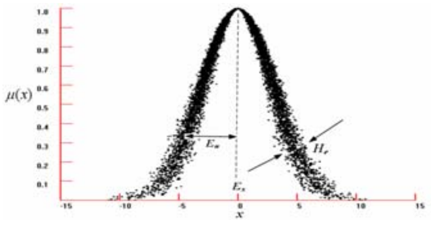


Figure 1 CG(0, 3, 0.3, 10000)

2.1.4 Backward Cloud Generator:

Backward cloud generator is a conversion model which can convert quantity numbers to a quality concept. It can convert the accurate data (x_1, x_2,…,x_n) to the quality cloud concept expressed by numerical characteristic (*Ex*,*En*,*He*). The algorithm of backward cloud generator can be summed as follows:

(1) According to the samples x_i, calculate its typical value $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, first-order absolute central moment of samples $\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{X}\right|$, and the variance of samples $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$;

(2) $Ex = \overline{X}$;

(3) $En = \sqrt{\frac{\pi}{2}} \times \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Ex\right|$;

(4) $He = \sqrt{S^2 - En^2}$.

2.1.5 Precondition Cloud Generator:

Given a universe *U*, in which the normal cloud model *C* can be expressed as *C* (*Ex, En, He*). If it is possible to generate the membership distribution of a specific point *a* which is in the universe *U* through cloud generator. Then the cloud model generator is called precondition cloud generator. It is illustrated in Figure 2.

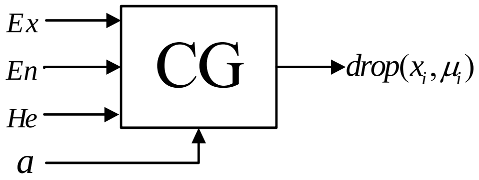


Figure 2. Precondition cloud generator

The algorithm of precondition cloud generator is as follows:

Input: The numeral characteristics of quality concept (*Ex, En, He*), the specific value *x* and the number of output cloud drops *n*.

Output: The cloud drops $(x_0,\mu_0),(x_1,\mu_1),\cdots(x_n,\mu_n)$

```
Begin
  For (i = 0； i < n； i++)
    Enni = Norm(En，He)
    $\mu_i = e^{\frac{-(x-Ex)^2}{2(Enni)^2}}$
    Return  drop(x， $\mu_i$)
  End For
End
```

The joint distribution of membership which generated by the specific value *x* and the precondition cloud generator is illustrated in Figure 3. All the cloud drops are located in the straight line of X = *x*.

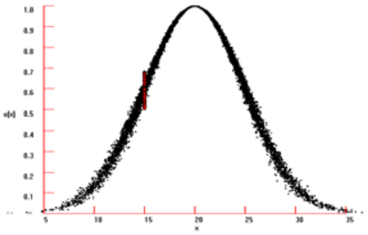


Figure 3 the joint distribution of membership

2.2 Hierarchical clustering based on cloud model

Hierarchical clustering method is an important kind of clustering methods. Its basic idea is to agglomerate or split the data recursively, and divide the data into nested class hierarchy or class pedigree. Such as CURE, BIRCH, CHAMELEON, etc. The advantages of theses methods lie in that they can basically discover the types of cluster of any shape, and get different size of the multi-level structure of clustering. But the algorithm's space complexity and time complexity is high, it's difficult to deal with large dataset. Therefore, actively explore new and more effective hierarchical clustering algorithm to achieve an effective large-scale datasets gathered on the clustering research is still an Open-Ended Question.

Agglomerative Hierarchical Clustering algorithms have high clustering quality, because they meet the characteristics of data. Currently, the traditional agglomerative hierarchical clustering algorithms often start from the individual data, set the separate data as a cluster, then through mergers with similar characteristics of the cluster to get a higher level of cluster. This will lead to agglomerative Hierarchical Clustering are more time complexity than other clustering algorithms. Thus, it was very early the hybrid clustering method is presented. Firstly, use other low time complexity of the cluster approach to get the initial dataset of clusters, then use the agglomerative hierarchical clustering approach to the initial cluster layers to the merger, and get on the end of the datasets. Initially, there are some hybrid clustering methods such as the hierarchical clustering methods based on partition, the hierarchical clustering methods based on grid. These hybrid clustering methods still have a problem, that their partition with datasets is absolute and don't take into account the fuzzy boundary problems of various clusters.

The paper proposes a novel hierarchical clustering algorithm based on the theory of cloud model. It uses a qualitative concept cloud model to express a cluster, and calculate all samples' membership degree matrix on cloud model to analyze their degree of membership of the clusters. Finally, according to Kosko subset of criteria to measure the various clusters of the merger, achieve the samples' multi-level clustering.

The basic steps of the method are as follows:

(1) Use the method of backward cloud generator on samples to divide the initial data, which gain the bottom cloud model that can describe the initial clusters.

(2) Use precondition cloud generator to calculate all samples' membership degree on cloud model, and get the samples' partition matrix.

(3) Traverse the partition matrix, and calculate the intersection element numbers of any two clusters

For a sample X_i, The more it approach the centre of a certain category X_i ($1 \le i \le c$，c for the number of the categories), the more its value for the category's membership degree function μ_{ik} on cloud model close to 1.When it located in the middle of two categories, the two membership degree functions are similar. Therefore, we can define the formulas as follows.

If the element x_k meet: ①Set the threshold ε_1, if $||\mu_{ik}-\mu_{jk}||<\varepsilon_1$, ②$\mu_{ik}=\max(\mu_{lk})$ $(l \in c)$ or $\mu_{jk}=\max(\mu_{lk})$ $(l \in c)$, then the element $x_k \in X_i \cap X_j$（$1 \le k \le n$, $1 \le i \le c$, $i \ne j$）.

(4) Define the threshold ε_2，using Kosko subset of criteria formulas to calculate f_k，if $f_k > \varepsilon 2$, then merge the clusters to satisfy the condition two by two. When it doesn't meet the conditions, go to the end. Or else continue to merge. In the course of merger, if cluster *A* comes from the merger of cluster *B* and *C,* the intersection element numbers of *A* with a cluster are the same as the intersection element numbers of *B* and C with the cluster.

Kosko subset of criteria formulas as follows:

$$f_K(A,B) = \begin{cases} 1, A = \emptyset \\ \frac{M(A \cap B)}{M(A)}, A \neq \emptyset \end{cases},$$

Where M (X) = | X |.
When judging two clusters with Kosko subset of criteria formulas, the smaller value of f_k(A, B), the more clear of their borders. The bigger value of *f_k(A, B)*, the more fuzzy of their borders. Then the two clusters can be merged when the value of *fk(A, B)* bigger than a certain value.

For the steps (3), the division matrix is a c×n matrix. So the time complexity for traversing the matrix is O(*n*). In step (4), the number of clusters less than or equal to constant C, the steps can be ended in limited time, so the time complexity for step(4) is a constant. Then the combined time complexity of the algorithm is O(*n*). Compared to the time complexity of traditional hierarchical clustering algorithm O($n^2\log_2 n$), this algorithm saves a lot of time.

3. METEOROLOGICAL SATELLITE IMAGE DATA MINING

3.1 Meteorological satellite image clustering experiments

The experiments collected more than 14000 district maps from FY-2C meteorological satellite between June and July in 2006, as well as six-hour rainfall data from weather stations during the same period as the original data sets.

The FY-2C satellite image has five bands, including four infrared bands and a visible band. Time resolution of the image is 30 minutes, which means that it can produce a cloud chart in every half hour. Between the neighbouring cloud charts there is some certain correlativity. Correlativity coefficient of two images with intervals of an hour is 0.9, and the value of the correlativity coefficient increases with the increase of time interval. Obviously, it is not appropriate for these strongly correlative cloud images to be used directly in cluster analysis, thus the result can not convey the essence of cluster analysis. What's more, for cloud images of visible light wave can not embody the main trait of the clouds at night, we need to resample the cloud chart data in the daytime. The time resolution of the rainfall data we have gathered is six hours, and the gathering time are respectively daily 2 o'clock, 8 o'clock, 14 o'clock and 20 o'clock, so we choose the daily 14 o'clock data as the sampling rainfall data. Because the rainfall climate is opposite in the cloud images characteristic has certain hysteresis quality, we choose the daily 13 o'clock 30 minutes data for the sampling cloud images. After such sampling rainfall data and the cloud images maintain the same time resolution and are well correlative. By spatial match between the meteorological station data and the cloud images, we obtain the FY-2C cloud images with five wave bands characteristic information which corresponds to various meteorological stations.

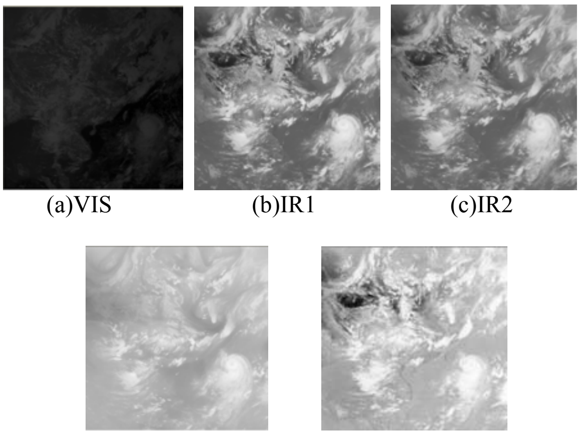
The whole 4000 3×3 pixels of FY-2C cloud images are chosen, containing six kinds of weather such as No rain, Light rain, Moderate rain Heavy rain, and Rainstorm; and the features of average bright degree of the cloud images are treated as sample data（sample number of each kind of weather is different）. With the method of hierarchical clustering based on cloud model, we cluster these sample data, and obtain the characteristic clusters of the luminance values of various wave bands which reflect different precipitation weather. It is illustrated in Table 1.

		VIS	IR1	IR2	IR3	IR4
No rain	Cluster centre	32	124	123	176	177
	Standard deviation	6.7	32.1	29.1	15.9	27.7
Light rain	Cluster centre	39	149	145	181	196
	Standard deviation	6.7	33.9	31.9	21.1	27.6
Moderate rain	Cluster centre	41	168	163	193	213
	Standard deviation	6.6	37.3	35.3	23.6	28.4
Heavy rain	Cluster centre	44	194	189	210	231
	Standard deviation	5.0	32.8	31.1	21.0	23.6
Rainstorm	Cluster centre	45	200	194	215	234
	Standard deviation	7.1	31.8	29.2	18.8	19.4
Heavy rainstorm	Cluster centre	47	205	197	216	237
	Standard deviation	2.8	22.0	23.2	18.4	13.8

Table 1 Clustering results

3.2 Examination of Weather classification

In order to check the precipitation weather classification results of spatial cluster model, we chose the satellite cloud images in July 6, 2006, 13 o'clock 30 minute as analysis objects which are shown in Figure 4(a)-(e). Through the construction of precondition cloud generator, the calculation of the cloud model's membership degree of each pixel on behalf of various precipitation types, we realize the classification of cloud images, and determine the types of precipitation weather. The result is illustrated in Figure 5.


(a)VIS (b)IR1 (c)IR2

(d)IR3 (e)IR4

Figure 4 Wave bands of FY 2C

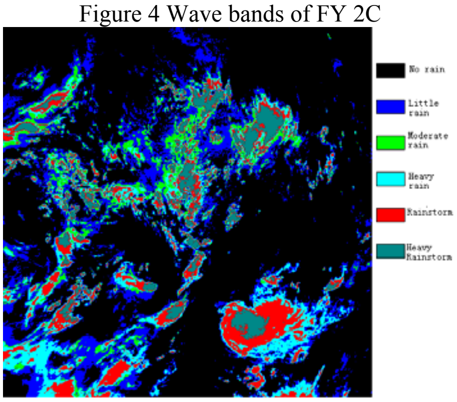


Figure 5 Results of precipitation weather classification

Compare the classification results of weather images with the actual values of precipitation in weather stations, the rates of accuracy are shown in Table 2.

	No rain	Light rain	Moderate rain	Heavy rain	Rainstorm	Heavy rainstorm
Number of actual stations	1754	425	65	27	15	5
Number of correction	1072	108	18	5	5	4
Rate of accuracy	0.61	0.25	0.28	0.19	0.33	0.80

Table 2 Analysis of accuracy rates

From the above table, it can be concluded that the accuracy rates of classification for the non-rain weather and heavy rainstorm weather are higher, but lower for other weather types. The causes may be that the cloud characteristics in non-rain weather and heavy rainstorm are obvious, and the corresponding brightness values in respective bands of FY-2C satellite cloud images are also concentrated. But for other weather types, the characteristics of cloud images are not obvious, and their brightness values are disperse, so the weather type discrimination based on satellite cloud images are difficult.

4. CONCLUSION

The application research of spatial data mining has become the future research direction of spatial data mining. The paper applied the method of hierarchy clustering based on cloud model to the relationship analysis of satellite cloud images and the precipitation weather. The experiments indicated that the proposed method is good for the discrimination of non-rain weather and heavy rainstorm weather, but not effective for other weather types. Our ongoing and further works include improving the hierarchical clustering based on cloud model and the accuracy rates of weather discrimination based on satellite cloud images.

ACKNOWLEDGEMENTS

This research is partially supported by National Key Basic Research and Development Program: 973 Program (2006CB701305; 2007CB311003), and by the Fund of the State Key Laboratory of Resources and Environmental Information System (A0682).

REFERENCES

Chen W.M., Yu F., 1994. Tentative study on estimating meiyu front precipitation with GMS numerical satellite data, *Journal of Nanjing Institute of Meteorology*, 17(1), pp.79-85.

Desbois,M.,Seze G., Szejwach G.,1982.Automatic classification of clouds on METEOSAT imagery application to high-level clouds, *J.Appl.Meteor*,(21),pp.401-402.

Hong M., Zhang R. Sun Z.B.,2006. A high dimension feature spaces clustering and corresponding weather classification for multi spectral satellite images, *Journal of Remote Sensing*, 10(2), pp.184-190.

Koffler R., Decotiis A.G., Rao P.K., 1973..A procedure for estimating cloud amount and height from satellite infrared radiation data. *Mon.Wea.Rev*, (101), pp.240-243.

Koperski K., Adhikary J., Han J., 1996. Spatial data mining: process and challenges survey paper, *SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June.

Li D.R., Cheng T., 1994. KDG-Knowledge discovery from GIS. *Proceedings of the Canadian Conference on GIS*, Ottawa, Canada, June 6-10:1001-1012

Li D.R., Wang S.L., Li D.Y., 2006. *The theories and application of spatial data mining*, science press.

Li D.Y., Du Y.,2007. *Artificial Intelligence with Uncertainty*, CRC Press, USA.

Li et al.,Maria C.V.R., Nelson J.F., Shi L.H., Leonardo D.A.., 2000. Rainfall estimation from meteorological satellite and radar data using multi-resolution wavelet transform and neural networks methods, *Journal of Nanjing Institute of Meteorology*, 23(2),pp.277-282.

Qin K., 2004. *Image data mining based on formal concept analysis*, Ph.D. Thesis, Wuhan University.

Shi C.X., Wu R.Z., Xiang X.K., 2001. Automatic segmentation of satellite image using hierarchical threshold and neural network, *Quarterly journal of Applied Metheorology*,12(1), pp.70-78.

Su F.Z., Zhou C.H., Lyne V., Du, Y.Y., Shi W.Z., 2004. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution, *ecological modelling*, (174), pp.421-431.

Wang Y.L., Zhang R., Sun Z.B., Niu S.J., Wang Q.L., Liang J.Y.,2005. Modification of cloud picture sample and automatic identification of cloud type based on fuzzy clustering method, *Advances in Marine Science*,23(2), pp.219-226.

Wang L.Z., Li J., Zhou F.X., The auto classification of 4-channels satellite images and its application in precipitation estimation, *J. Atmo Sci Sinica*, 22(3),pp.371-378.

Welch,R.M., Navar M.S., Sengupta S.K.,1989. The effect of resolution upon texture-based cloud field classification, *J.Geophys.Res*, (94), pp.14767-14781.

Yang Y.B., Lin H., Guo Z.Y., Jiang J.X.,2007. A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis, *Computers Geosciences*, (33), pp.20-30.

Yu F.,1998. *The characteristic extraction from GMS multi-spectrum satellite images and its application in mid-scale numerical forecast model*, Ph.D. thesis, Nanjing university.