Research of the Multi-Dimensional Cloud Classification Algorithm

Li Qin and Bing Li

***Abstract*—Data classification is the basic approach of data mining and Knowledge discovery in databases (KDD). In recent years, cloud classifier based on the cloud theory has been proposed. The most difference between cloud classifier and the traditional classifiers was that classified boundary of cloud classifier is fuzzy. Since current research only focus on the one-dimensional cloud generator algorithm, so this paper presents the classification algorithms based on the multi-dimensional cloud generator. Moreover, to resolve the complexity of classification which was brought by multi-dimension independent samples, the author proposes a method to so1ve the dimensionality reduction problem of multi-dimensional samples by one-dimensional cloud charts. Finally, the accuracy of cloud classifier is verified by a classification experiment on a texture database.**

***Index Terms*—Cloud model, classification algorithm, dimensionality reduction**

I. Introduction

Data classification is the categorization of data for its most effective and efficient use, so it is the basic approach of data mining and Knowledge discovery in databases (KDD). In recent research , the most widely used classifiers are the neural network , support vector machines, k-nearest neighbours, decision tree and RBF classifiers . In recent years, some scholars build cloud classifier with cloud model[1,2,3]. Cloud model which was introduced by Li Deyi[4] is an effective tool in transforming between qualitative concepts and their quantitative expressions which was built on the basis of fuzzy logic and probability theory. Different from the traditional classifiers, the classified boundary of the cloud classifiers is fuzzy, but the trend of classification is stable, so it is more close to the distribution of actual data and the thinking way of human. Therefore, this paper focuses on the study of the cloud classification algorithm. The present research only used the one-dimensional cloud generator algorithm, not considered multi-dimensional cloud generator algorithm. So this paper proposes the classification algorithms based on the multi-dimensional cloud generator, and in order to measure the effect of the features in the classification, this paper propose a method to reflect the feature differences between the categories by one-dimensional cloud charts, and so to so1ve the dimensionality reduction problem for multi-dimensional sample. Finally, effectiveness of the cloud classifier proposed in this paper is verified by a classification experiment on a texture database.

Manuscript received February 28, 2012; revised April 29, 2012.

L. Qin and B. Li is with State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China.

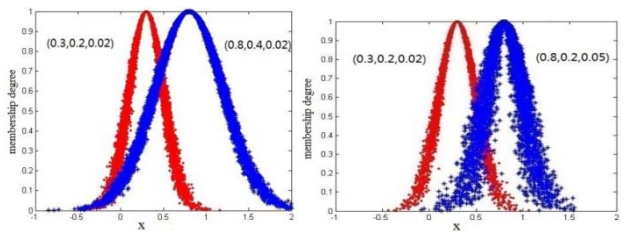
L. Qin is with School of Science, Huazhong Agricultural University, Wuhan 430070, China (e-mail: qinli0606@126.com).

B. Li is with School of Computer, Wuhan University, Wuhan 430072, China.

II. Cloud Model

Cloud model is more applicable and universal in the representation of uncertain notions. The transforms are performed between a qualitative concept and a quantitative representation by two cloud generators: one-dimensional forward cloud generator and one-dimensional backward cloud generator, and both generators can be extended to multi-dimensional[5].

In cloud model theory, the fuzziness and randomness of an uncertain concept are merging together in the three numerical characteristics: expected value, entropy, and hyper entropy which can be expressed as (Ex, En, He). Ex is the point in the universe of discourse that could represent this qualitative concept properly. En is the uncertainty measurement of the qualitative concept. it can reflects accepted range of number fields of the concept on the universe of discourse, just as Fig. 1(a) shows. He is the uncertain degree of Entropy. Generally, the greater is the Hyper Entropy, the greater is the dispersion degree of cloud drops, the greater is the randomness of the membership degree. Just as Fig. 1 (b) shows.



(a)The change of cloud in different Entropy (b) The change of cloud in different Hyper Entropy

Fig. 1. One-dimensional cloud in different numerical characteristics

III. Cloud Classifier

The common process of classification can be described as two steps, one is sample training the other is classification. The approach of cloud classifier is similar, which is depicted in Fig. 2.

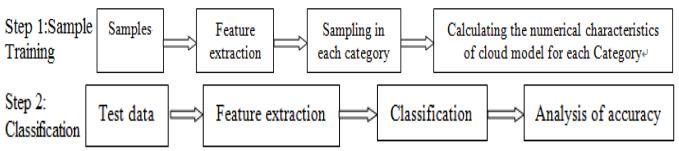


Fig. 2. The approach of cloud classifier

Step 1: The first work of sample training is extracting features from the sample set, and the features which are extracted constructs a matrix T; The next work is selecting the same number of random samples from the matrix T for each Category, each samples build some new matrices $T_i^{'}$ (i is the category number); Finally, set feature $P_{i,j}$ (j is the feature number) in the matrix $T_i^{'}$ as a group of cloud drops, and calculate the numerical characteristics of cloud model with Multi-dimensional backward cloud algorithm. This paper employ the backward cloud algorithm with no membership degree[6], the specific algorithm is as follows:

Input: Feature $P_{i,j}$, and cloud drops $x_{i,j,l}$ in $P_{i,j}$ (i=1..N; j =1..M; l =1..W; l is the sample number).

Output: The numerical characteristics ($Ex_{i,j}$, $En_{i,j}$, $He_{i,j}$) of $P_{i,j}$ ((i=1,2,…N; j =1,2,…M).

(1) For i =1 to N

(2) For j = 1 to M

(3) Compute the mean of the cloud drops $x_{i,j,l}$, $\overline{X} = \frac{1}{W}\sum_{l=1}^{W} x_{i,j,l}$

(4) $Ex_{i,j} = \overline{X}$

(5) $En_{i,j} = \sqrt{\frac{\pi}{2}} \times \frac{1}{W}\sum_{l=1}^{W}\left|x_{i,j,l} - Ex_{i,j}\right|$

(6) Calculate the variance of the cloud drops $x_{i,j,l}$, $S^2 = \frac{1}{W-1}\sum_{l=1}^{W}\left(x_{i,j,l} - Ex_{i,j}\right)^2$

(7) $He_{i,j} = \sqrt{S^2 - {En_{i,j}}^2}$

(8) END FOR j

(9) END FOR i

Step 2: Before classification, we also need to extract features from test data, after that multi-dimensional forward cloud classification algorithm was exploited to calculate the membership degree of each data for each category. At last, the category of every data belongs to will be judged by the maximum value of the membership degree, the specific algorithm is as follows:

Input: The numerical characteristics ($Ex_{i,j}$, $En_{i,j}$, $He_{i,j}$) output from last step, and weights k_j of $P_{i,j}$, and new samples $y_{l,j}$ (i =1,2,…N; j =1,2,…M; l =1,2,…W$^{'}$).

Output: the category number of new samples F_l (l =1,2,…W$^{'}$).

(1) For l=1 to W$^{'}$

(2) For i =1 to N

(3) For j = 1 to M

(4) Generate normal random number $En'_{i,j}$ based on $En_{i,j}$ as expected value and $\left(He_{i,j}\right)^2$ as variance

(5) $x_i = x_i + k_j \times \left(-\frac{\left(y_{l,j} - Ex_{i,j}\right)^2}{2\left(En'_{i,j}\right)^2}\right)$

(6) END FOR j

(7) Count the membership degree $u_i = e^{x_i}$

(8) END FOR i

(9) Find out the maximum value in u_i, and set the number i of it to F_l, that is , $F_l = i$

(10) END FOR l

The weights k_j in the algorithm will be defined in the next section. While the analysis of accuracy is the last work of classifier, the specific contents will be described in the experiment.

IV. Dimensionality Reduction

Reflecting the most of the feature information of the original dataset with fewer features is the goal pursued by dimensionality reduction, the traditional dimensionality reduction algorithms are PCA, ICA, etc. How to measure the effect of a feature in the classification? The author believes that it must be able to fully reflect the differences between the categories, and the difference between the categories is the bigger the better. So this paper consider "the bigger difference" as a dimensionality reduction standard, and proposes a method to reflect the feature differences between the categories by one-dimensional cloud charts, so as to so1ve the dimensionality reduction problem for multi-dimensional sample.

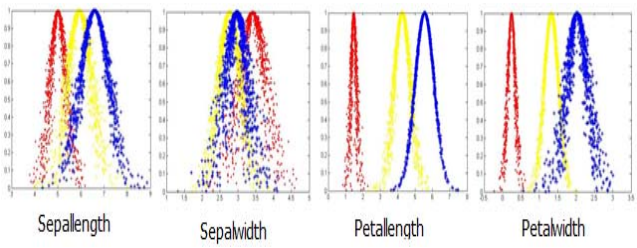


Fig. 3. One-dimensional cloud chart of each feature of iris data set (note: sepallength, sepalwidth petallength and petalwidth are the entire property name)

Take iris data set for example, iris data set has 3 classes and 4 properties. And give the one-dimensional normal cloud chart of each feature of iris data set, which are shown in Fig. 3. Each chart consists of three clouds, representing the three categories. It can be found that the three clouds in the chart of sepallength and sepalwidth are more concentrated and discrete, this shows that the expected value of those clouds are closer and Hyper Entropy are also relatively large. These indicate that the value of sepallength and sepalwidth between the categories is relatively close in the overall and the category of the samples belong to are rather vague, in other words, the value difference of sepallength and sepalwidth between the categories is small, and not fit for the classification, in contrast, petallength and petalwidth conform standard of "the bigger difference", so benefit to the classification.

After the analysis above, the feature differences between the categories are closely related to expected value and hyper entropy, so the author proposes a new formula to compute weights of the features in the classification. Now set Multi-dimensional cloud numerical characteristics as ($Ex_{i,j}$, $En_{i,j}$, $He_{i,j}$), let i =1,2,…N; j =1,2,…M , then the mean of expected value is :

$$\overline{Ex_j} = \frac{\sum_{i=1}^{N} Ex_{i,j}}{N} \tag{1}$$

Variance of expected value is:

$$s_j^{\,2} = \frac{\sum_{i=1}^{N}\left(Ex_{i,j} - \overline{Ex_j}\right)^2}{N-1} \tag{2}$$

The formula is defined as follow:

$$Wt_j = \frac{s_j}{\overline{Ex_j}} \times \sum_{i=1}^{N} \frac{Ex_{i,j}}{He_{i,j}} \tag{3}$$

According to the formula, the calculation results are shown in descending order in Table.1. Because weights of Sepallength and Sepalwidth is too small, so they can be removed from the process of classification, and the other two features will be used in the classification with the weights shown in Table I.

TABLE I: THE EFFECT OF THE FEATURES IN THE CLASSIFICATION

Feature	Wt_j	Feature	Wt_j	Feature	Wt_j	Feature	Wt_j
Petallength	0.6142	Petalwidth	0.2692	Sepallength	0.0803	Sepalwidth	0.0364

At last, in order to verify the accuracy of classification of cloud classifier, the iris data set was classified with several different classification methods, the accuracy of them are shown in Fig.4. In this experiment, the iris data was reduced through PCA algorithm, besides the cloud classifier. With comparison, the cloud classifier has a good effect.

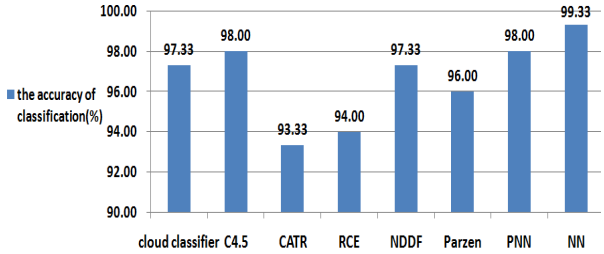


Fig. 4 . Comparison of the accuracy rate of classification(note: programming language of cloud classifier is matlab, and other algorithms used in comparative experiments is come from the classification toolbox [9] in MATLAB.)

V. EXPERIMENT

In this experiment, the author exploit cloud classifier on the classification of texture images, The texture database was supplied by Svetlana Lazebnik[8], etc. The texture database features 25 texture classes, 40 samples each. All images are in gray scale JPG format, 640x480 pixels. The approach of experiment is as the following steps:

TABLE II: THE EFFECT OF THE FEATURES IN THE CLASSIFICATION

Feature	Wt_j	Feature	Wt_j	Feature	Wt_j
Sum Variance	0.2361	angular second Moment	0.0886	contrast	0.0352
sum entropy	0.1633	variance	0.0692	inverse difference moment	0.0269
entropy	0.1081	sum average	0.0588	difference entropy	0.0242
correlation	0.1079	difference Variance	0.0588	difference average	0.0227

(1)First, extract the features of texture image, by calculating the Ggray-level Co-occurrence Matrix of gray scale image, twelve eigenvalues are extracted. Those eigenvalues build a matrix T.

(2)Then, calculate numerical characteristics of each feature by sample training method.

(3) According to the numerical characteristics and the formula proposed in former section, the weight of each feature can be computed which are sorted in descending order in Table.2. The first four features were selected in cloud classification and the accuracy of it is shown in Fig.5.

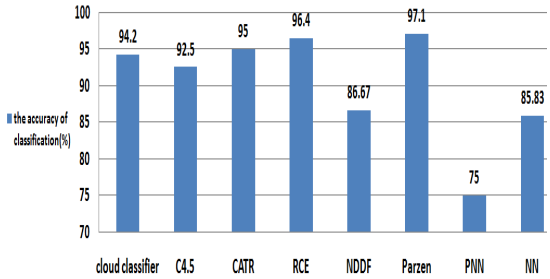


Fig. 5. Comparison of the accuracy rate of texture classification

By comparing the accuracy of classification of cloud classifier and other classifier, which are also shown in Fig.5, the author find the result is quite satisfactory.

VI. CONCLUSION

Data classification is the process of separation in which data is divided into several groups, while cloud model is an important uncertain theory in artificial intelligence. Classified boundary of cloud classifier is fuzzy, but the trend of classification is stable, so cloud classifier is more close to the thinking way of human. This paper mainly discusses the study of the algorithm design of multi-dimensional cloud classifier, and proposes a method to so1ve the dimensionality reduction problem of multi-dimensional samples by one-dimensional cloud model. Finally, we have a good result in the classification experiment.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under grant No. 60970017, 60903034, 61100017, 61100018, 61170025, 61170026； the Natural Science Foundation of Hubei Province of China under Grant No. 2011CDB055, 2011CDB069, 2010CDB08503.

REFERENCES

[1] D. Liu, G. Y. Zhang, G. Z. Liu, and Y. Sha, "A Cloud Classification Model Based on Feature Selection," *Journal of Beijing Institute of Petro chemical Technolog*, vol.10. no.1, pp.10-14, 2011.

[2] E. X. Pi, H. F. Lu, B. Jiang, J. Huang, Q. F. Peng, and X. Y. Lin, "Precise plant classification within genus level based on simulated annealing aided cloud classifier," *Expert Systems with Application,* vol. 38, pp. 3009-3014. 2011.

[3] H. F. Lu, E. X. Pi, Q. F. Peng, L. L. Wang, and C. J. Zhang, "A particle swarm optimization-aided fuzzy cloud classifier applied for plant numerical taxonomy based on attribute similarity," *Expert Systems with Applications*, vol. 36, pp. 9388-9397. 2009.

[4] D.Y. Li, C. Y. Liu, and Y. Du, etc. "Artificial Intelligence with Uncertainty," *Journal of Software*, vol. 15, no. 11, pp. 1583-1594, 2004.

[5] D. Y. Li and C.Y. Liu, "Study on the Universality of the Normal Cloud Model," *Engineering Science*, vol. 6, no. 8, pp. 28-33. 2004.

[6] C. Y. Liu, M. Feng, X. J. DAI, and D. Y. Li, "A New Algorithm of Backward Cloud," *Journal of System a Simulation*, vol. 16, no. 11, pp. 2417-2420. 2004.

[7] G. D. Stork and E. Y. Tov, *Computer Manual in MATLAB to accompany Pattern Classification, 2nd Edition*, 2004, pp. 978-0-471-42977-7.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "A Sparse Texture Representation Using Local Affine Regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1265-1278. 2005.