Cloud Model-based Data Attributes Reduction for Clustering

XU Ru-zhi, NIE Pei-yao, LIN Pei-guang, CHU Dong-sheng
School of Information Engineering, Shandong University of Finance, Jinan 250014, P.R. China
E-mail:rzxu@sdfi.edu.cn

Abstract

Data reduction, which can simplify large scale data and not lose useful information, is an important topic of knowleadge discorvery, data clustering and classification. Aiming to solve the current problem that continuous attribute in algorithm of clustering or classification has to be discrete, a new algorithm of data reduction based on cloud model is put forward. By use of cloud model, this algorithm calculates each conditional attribute's importance to decision-making attribute(s), and obtains the reduction attributes by virtue of greedy algorithm. This new data reduction algorithm was verified by some experiments and was proved to be stable and efficient.

1. Introduction

Real world data sets usually have many features; some of them are irrelevant or redundant[1]. The exist of such feature will increase the complexity of data mining task, add noise to the data, and make the result of data mining difficult to understand[2]. For data mining, a successful choice of features can improve the accuracy, save the computation time and memory space, and simplify its results. Feature selection is a process that chooses an optimal feature subset according to a certain criterion [3]. It will reduce the dimensionality of the data and may allow data mining to operate faster and more effectively. In some cases, accuracy on data mining results can be improved; in others, the result is a more compact, easily interpreted representation of the target concept [5].

Data reduction is an important concept while Rough Set applies on data analyze. But before Rough Set processes the data, it demands that the data must be discrete but this would lose more useful information. Taking advantage of the cloud theory, this paper proposed a new data reduction algorithm aiming at the continuous attributes, which can reduce the continuous attributes directly and does not need to be discretized[5].

In this paper, we present a new data reduction algorithm based on Cloud Model. This algorithm can process the continuous attributes directly and does not need to discrete them by the virtue of greedy algorithm. Compared with some other algorithms, not only is this one simply and efficient but also the experiment results are very perfect.

2. Cloud Model

The Cloud Theory is a new one which is presented by Prof. De-yi Li based on Fuzzy Logic and Probability and is used to processing the uncertainty and providing a means of both qualitative and quantitative characterization of linguistic atoms [6].

Linguistic Atoms: We imagine a linguistic variable that is semantically associated with a list of all the linguistic terms within a universe of discourse. Each possible value of a linguistic variable represents a fuzzy concept, and it is defined as a linguistic atom. In the more general cased, a linguistic variable is a quintuple:

$$\{X, T(x), U, S, C_X(u)\} \quad \text{Eq.1}$$

X is the name of the variable. *T(x)* is the term-set of *X*; that is, the collection of its linguistic atoms. *U* is a universe of discourse. *S* is a syntactic generator which generates the atoms in *T(x)* and *U*. More precisely, the compatibility function maps the universe of discourse into the interval [0,1] for each *u* in *U*.

It is important to understand the notion of compatibility functions. Consider a set of linguistic atoms *T* in a universe of discourse *U*. For example, the linguistic atom "warm" is the interval U=[30° F,100° F]. *T* is characterized by its compatibility function $C_X : u \in [0,1]$. The statement that the compatibility of, say 68° F, with "warm" is 0.3, has a relationship both to fuzzy logic and probability.

Normal Cloud Model: Normal Cloud Models are the most fundamental and useful in representing lin-

e-Forensics 2008, January 21-23, 2008, Adelaide, Australia.

guistic atoms. We could also use the normal membership function to represent the mathematical expected curve (MEC) of the cloud. The digital parameters of a normal Cloud Model characterize the quantitative meaning of a linguistic atom. Gaussian kernels are used in a very effective way in characterizing the normal membership clouds.

EXPECTED VALUE Ex. The expected value Ex of a normal Cloud Model is the position at the universe of discourse, corresponding to the center of gravity of the cloud. In other words, the element Ex in the universe of discourse fully belongs to the linguistic atom represented by the Cloud Model. It is very easy to determine Ex in practical applications.

ENTROPY En. The entropy is a measure of the fuzziness of the concept over the universe of discourse showing how many elements in the universe of discourse could be accepted as the linguistic atom. The entropy is a generic notion, and it need not be probabilistic. The entropy decreases as the MEC bandwidth decreases, only if upon the narrowing membership cloud turns to be a precise numerical value is formed, the entropy becomes zero. The mathematical expected function of the normal membership cloud with its expected value Ex and entropy En may be written as:

$$MEC_A(u) = e^{-(u-Ex)^2/2En^2} \quad \text{Eq.2}$$

DEVIATION He. Looking at the normal Cloud Model in detail, in Figure 1, we see that its thickness is uneven. The deviation *He* is a measure of the randomness of membership function. Close to the waist, corresponding to the center of gravity $(Ex, \sqrt{2}/4)$, the degree of membership is most dispersed, while at the top and bottom the focusing is much better. Therefore, the maximum deviation *He* really comes from the randomness of the membership degree at the waist part of the cloud.

Forward Cloud Generators*:* Given three digital characteristics Ex, En, and D, to represent a linguistic atom, say "youth", the generator could produce as many drops of the cloud as you like. Figure 2 shows the shapes with 3000 drops generated respectively with the parameters Ex=25, En=3, and He=0.1.

Cloud-drops may be generated on conditions. Figure 3a shows a generator producing drops under a given numerical value *u* in the universe of discourse, *U*; while Figure 3b shows the same generator under the condition of a given membership degree μ. All the drops generated in Figure 3a have the same value of u in the universe of discourse, and with different Gaussian distributed membership degrees μ_i; whereas all the drops generated in Figure 3b have the same membership degree μ, and with different Gaussian distributed numerical value u_i in the universe of discourse.

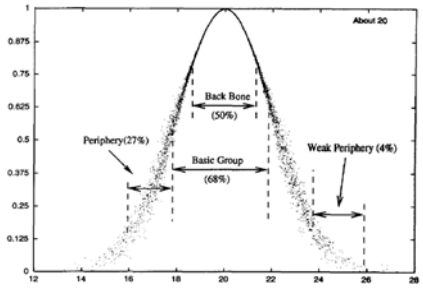

Fig.1 Contributions with different groups

Backward Cloud Generators: Given a limited set of drops, $drop_i(u_i, \mu_i)$, as samples of a normal Cloud Model, the three digital characteristics Ex, En, and D could be produced to represent the corresponding linguistic atom. This kind of cloud generators are called backward cloud generators, denoted by G^{-1}.

Since all linguistic atoms are represented by the cloud model, the forward and backward cloud generators can be served interchangeably to bridge the gap between quantitative and qualitative.

Definition 1 Collision Object of Attribute (COA): Given the Decision Table S=(U,C∪D), for any conditional attribute c∈C, the symbol u_i^c denotes the value of the i^{th} object about attribute c. If the value of decision attribute(s) can be decided by u_i^c, we define it No-Collision, otherwise collision.

Definition 2 Importance of Attribute (IOA): Given the Decision Table S=(U,C∪D), for any conditional attribute c∈C, if the number of No-Collision about the attribute c is N, we define IOA is $Imp_c = N/|U|$.

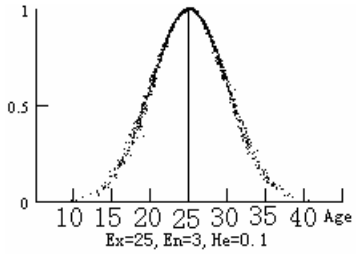

Fig.2 a Normal Cloud with 3000 drops

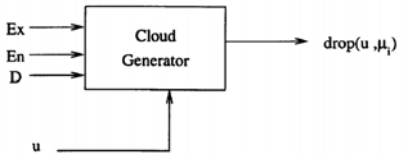

Fig.3 (a) On the condition of "u"

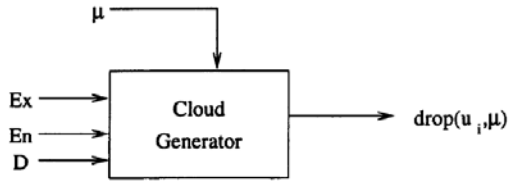

Fig.3(b) On the condition of "μ"
Fig.3 Generators on Condition

3. Cloud Model-Based Data Reduction

Cloud Model can carry out the transformation between qualitative and quantitative characterization of linguistic atoms. Using the backward cloud generator based on X-information, we can obtain the qualitative concept of linguistic atoms from the quantitative data; using the forward generator, we can get the quantitative data from the qualitative concept of linguistic atoms.

Given a group of data, we can get their digital characterizes firstly by using the backward cloud generator based on X-information; Then making use of the Entropy or forward cloud generator, we can get each data's contribution (importance) to its attribute. By this way, we can compute all the collision data in each attribute and then we can get the importance of each attribute. After getting the importance of each attribute, we can get the deduct attributes by the thought of greedy.

3.1. Importance of the attributes based on Cloud Model

In order to compute Impc, we must get the number of COA. By the definition, we give the following rules which can judge a record is COA or not. Firstly, we give the rule 1 which can an element is subjected to a cloud or not; and then by rule 1, rule 2 give a method to decide a record is COA or not.

Rule 1 Given element x0, threshold of contribution a, three characterizes of normal cloud model, by use of the generator described by Figure 3a, we can get the contribution a' of x0, if $a' < a$, we define x0 is belong to the cloud, otherwise not.

Rule 2 For any conditional attribute and an element x0, we first classified the values of the conditional attribute by the decision attribute(s) and then get the clouds of each classification. By Rule 1, if the element x0 is only belong to only one cloud, we define that the element x0 is not a COA, otherwise it is.

So, we give the following algorithm to compute IOA:

Algorithm 1 the Computation of IOA of conditional attribute
Input: The Decision Table: S=(U,C∪D), threshold: a
Output: IOA:Imp_c
Step 1: Classify the values of each conditional attribute by decision attribute(s)
By use of the forward cloud generator based on X-information, get each classification's cloud
Step 2: Initialize the counter of No-Collision: countRecCollision = 0
Step 3: For each collision object
By Rule 1, count how many clouds does it subject to. Mark the number with N
If N = 1 then
Mark the object with No-Collision
countRecCollision = countRecCollision + 1
Step 4: Compute the IOA: Imp_c=countRecCollision/|U|

3.2. Data Reduction algorithm Based on Cloud Model

Algorithm 2 Data Reduction algorithm Based on Cloud Model
Input: The Decision Table: S=(U,C∪D), threshold: a
Output: The deduction attributes collection
Step 1: For each conditional attribute
Classify the values of each conditional attribute by decision attribute(s)
By use of the forward cloud generator based on X-information, get each classification's cloud
Step 2: Initialize the mark collection of all Collision object and record the clouds
Step 3: If the mark collection is not empty, then go to step 4
Step 4: If all conditional attributes have been merged into deduction attributes collection, then stop
Step 5: For each conditional attributes who has not merged into deduction attributes collection
Compute the IOA by use of algorithm 1
And mark the collision
Step 6: Choice the attribute whose IOA is maximum and merge it to the deduction attributes collection
Step 7: Go to step 3

3.3 Time Complexity Analyze

Supported that the number of the decision attribute(s)' value was *countDec*, and the number of the conditional attributes is *countCon*. In algorithm 1, the step1 can be finished in linear time. In step 3, the algorithm need to traverse every cloud and record, so its complexity is O(countDec * |U|).

The algorithm 2 spends time mainly on step 5. In the worst case, no conditional attributes can be deduced, the time complexity is O(*countCon* * *countCon* * *countDec* * |U|). In the best case, only on conditional attribute resolved all the collision, then the time complexity is O(*countCon* * *countDec* * |U|). It is obvious that this time complexity is much less than [7] and [4]. Furthermore, this new algorithm does not need to discrete the continuous attributes, as is a distinguishing feature of this algorithm.

4. Experiments and Analyzes

In order to evaluate the performance of our algorithm, we have tested it on UCI repository of machine learning database [8]. We also use RSES Exhaustive reducer of Rosetta to generate all reducts of data set, and compare all reducts with our result. The experiment results are list at table 1.

In table 1, first column is the name of data set, and second one is attribute number, the third is instance number. Numbers of attributes selected by our algorithm are list in the fourth column. In fifth column, yes means result is the shortest reduct of data set.

Table 1. Experiment Result

Data set	Attributes	Instances	Selected Attributes	Optimal
Breast Cancer	11	699	4	Yes
Bridges	13	108	4	No
Tic-tac-toe	10	958	7	Yes
House-votes	17	435	8	Yes
Zoo	17	101	5	Yes

As the experiment result show, our algorithm can find out shortest reduct in most times. Even shortest reduct cannot be found; our algorithm can generate a shorter reduct without irrelevant and redundant attributes. According complexity analysis in section 4, our algorithm is faster than algorithms of [7] and [9].

5. Conclusion

In this paper, we present a new data reduction algorithm based on Cloud Model. This algorithm can process the continuous attributes directly and does not need to discrete them. The algorithm is simple but very efficient but the algorithm analyze indicates that its time complexity is very low. The experiment shows that this algorithm can carry out the task of data deduction and the results are very ideal and can satisfy the need of application.

Acknowledgement

This work is supported by Natural Science Foundation of ShanDong Province, P.R. China (A200621), and by the Education Foundation for S&T Development of ShanDong Provine, P.R. China (J06P07).

References

[1] Constantinopoulos C, Titsias M K, Likas A. Bayesian feature and model selection for Gaussianmixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(6):1013-1018.

[2] Griffiths T L, Ghahramani Z. Infinite latent feature models and the Indian buffet process[C] // Proc. Of the Advancesin Neural Information Processing Systems. Vancouver, Canada, 2005, 18: 475-482.

[3] Yang Chun-Mei, Wan Bai-Kun, Gao Xiao-Feng. Selections of data preprocessing methods and similarity metrics for gene cluster analysis. Progress in Natural Science, 2006, 16(6):607-613.

[4] Meeds E, Ghahramani Z, Neal R, Sam R. Modeling dyadic data with binary latent factors // Proc. Of the 11th International Conference on Artificial Intelligence and Staticstics. San Juan, Puerto Rico, 2007.

[5] Hall, M.A. Correlation-based Feature Selection for Machine Learning. PHD thesis. Department of Computer Science, University of Waikato, Hamilton (1999).

[6] Li Deyi, Cheung DW, Shi Xuemei et al. Uncertainty Reasoning Based on Cloud Models in Controllers. Computers and mathematics with Application, Elsevier Science, 1998,35(3):99-123

[7] Jing Zhang, Jianmin Wang, Deyi Li, et al. A New Heuristic Reduct Algorithm Base on Rough Sets Theory. LNCS 2762, pp. 247–253, 2003. Berlin :Springer-Verlag

[8] Blake, C. L., Merz, C. J.:UCI Repository of Machine Learning Databases.
http://www.ics.uci.edu/mlearn/MLReposityory.html

[9] Keyun Hu, lili Diao and Chunyi Shi: A Heuristic Optimal Reduct algorithm. 22nd Intl. Sym. on Intelligent Data Engineering and Automated Learning (IDEAL2000), Hong Kong, (2002)