Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

Continuous Attribute Reduction Method Based on an Automatic Clustering Algorithm and Decision Entropy

Hairong Sun[1], Rui Wang[1,2], Bixia Xie[1,2], Yao Tian[1,2]

1. School of Control and Computer Engineering，North China Electric Power University，Baoding 071003
E-mail: esunhair@163.com

2. Hebei Engineering Research Center of Simulation & Optimized Control for Power Generation，Baoding 071003
E-mail: wangrui_621@126.com, xiebixiaccc@163.com, 2425436588@qq.com

Abstract: The data in the process of power plant are mostly continuous. Aimed at the problems existing in the discretization of the continuous attributes with associated relationship, a strategy combining k-means clustering algorithm and the degree of compatibility of the decision table was proposed. The degree of compatibility was taken as judgment conditions to determine whether a new cluster type is added, while ensuring that the degree of compatibility can be accepted. This method take full account of the relationship among variables. The automatic clustering is realized when discretizing the continuous attributes, which makes the clustering results effective and reasonable. An attribute reduction algorithm based on approximate decision entropy is used to extract the main relations of the system, and the dimensionality of the data set can be effectively reduced. The improved scheme is applied to continuous data of main steam temperature system of thermal power plant, and the validity of the method is verified.

Key Words: Automatic clustering, Continuous attribute, Attribute reduction, Decision Entropy, Degree of compatibility

1 Introduction

With the rapid development of intelligent and the rapid growth of data, how to analysis the massive correlate data and make knowledge extraction has been widely concerned in recent years. Rough set theory is a mathematical tool for dealing with uncertain information, which was proposed by the Polish scientist Pawlak in 1982. Attribute reduction is the core of rough set theory. It can eliminate the redundant attribute and reduce the dimension of data set under the condition of keeping the classification ability unchanged[1]. It is extensively studied and applied in the fields of large data processing and data mining[2][3].

The data which Rough set theory can deal with is discrete data, but the process data are mostly continuous in the power plant. Therefore, the discretization of continuous data is the premise of applying rough set theory. Discretization methods include global discretization and local discretization, supervised discretization and unsupervised discretization [4][5]. Clustering analysis is a typical unsupervised discretization method, which extracts the similarity of objects by feature data and then classifies them, which can identify the regularity from the complicated data. K-means clustering algorithm is an important clustering algorithm, with the advantages of simple, easy implementation and fast convergence. Traditional k-means clustering algorithm has two shortcomings, one is the initial value of clustering is chosen randomly, and the other is the number of clusters must be specified in advance. Aimed at the first shortcoming, an algorithm is proposed to optimize the initial cluster center by using the center of gravity in reference [4]. In reference [5], the initial clustering center with higher quality is selected by weighting function. And the genetic algorithm is used to solve the initial cluster center in reference [6]. Aimed at the second shortcoming, an algorithm based on recursion is proposed in reference [7]. These improved algorithms can achieve good clustering effect，but for the systems in which variables are associated , the application effect is not very well.

In this paper, a strategy combining k-means clustering algorithm and the degree of compatibility of the decision table was proposed. The degree of compatibility was taken as Judgment conditions to determine whether a new cluster type is added, while ensuring that the degree of compatibility can be accepted. This method take full account of the relationship among variables. When discretizing the continuous attributes with associated relationship, the automatic clustering is realized, which makes the clustering results more reasonable. An attribute reduction algorithm based on approximate decision entropy is used to extract the main relations of the system, and the dimensionality of the data set can be effectively reduced. The improved scheme is applied to continuous data of main steam temperature system of thermal power plant, and the validity of the method is verified.

2 Concept Description

Rough set theory is a mathematical tool dealing with imprecise and uncertain information. It doesn't need to provide the previous verify information except the data set. Dealing with the uncertainty is objective[8]. The basic definitions are as follows[8]:

Definition 1 Given a knowledge base K=(U,S), where U is the universe of discourse, S is an equivalence relation on U, and the equivalence relations are $P\subseteq S, \forall R\in P$. If there exists IND($P$)=IND($P$-{$R$}), then knowledge R is known as unnecessary in P, otherwise it is necessary. If for every $R\in P$, R is all necessary, then P is independent, otherwise is not independent. Attribute reduction is to delete the irrelevant redundant attribute while keeping the knowledge base classification ability unchanged.

*This work is supported by the Fundamental Research Funds for the Central Universities (2016MS143)

With the decision table $DT=(U,C \cap D,V,f)$, where U is the universe of discourse, C is the conditional attribute set, D is the decision attribute set, V is the range of the information function f, and f is information function of the decision table. Let $X \in U/IND(C)$, $Y \in U/IND(D)$, $\forall x \in X$, the description of equivalence class X is

$$des(X) = \underset{\forall \alpha \in C}{\wedge} (\alpha, \alpha(x))$$

$\forall y \in Y$, the description of equivalence class Y is

$$des(Y) = \underset{\forall \beta \in D}{\wedge} (\beta, \beta(x))$$

Then we have the Definition 2 and 3 as follows.

Definition 2 The decision rule from X to Y is

r: $des(X) \rightarrow des(Y)$, $Y \cap X \neq \varnothing$

Definition 3 The degree of certainty of decision rules is

$$\mu(X,Y) = \frac{|Y \cap X|}{|X|}$$

The degree of certainty of decision rule reflects the adequacy of $des(X) \rightarrow des(Y)$.

Definition 4 The degree of compatibility of decision table [9] is

$$T = \min_{i=1}^{t} \{M_{X_i}\}$$

$$M_{X_i} = \sup_{j=1}^{r} \{\mu(X_i, Y_j)\}$$

Where t is the equivalence class of condition attribute C and r is the equivalence class of decision attribute D. T is the degree of certainty of the rule with the smallest degree of certainty in the rule set, which is taken as the degree of compatibility of decision table[9].

3 The Method of Discretization of the Continuous Attributes with Associated Relationship

As the rough set can only deal with discrete data, as for continuous data, it should be discretized at first. In this paper, k-means clustering method is used to discretize continuous data, and the degree of certainty of decision table is introduced as the feedback information of discrete result, considering the inherent relationship among variables, the decision table is discretized as a whole. It is necessary to normalize the data samples before discretization, and the data range is between [0,1].

3.1 The Improved K-means Clustering Algorithm

The traditional k-means clustering algorithm usually selects a sample point randomly as the first clustering center. The clustering results are much sensitive to the initial clustering center and the clustering results are unstable. In this paper, the improved k-means clustering algorithm are adopted and the initial clustering center is a certain value, the average of all data points. Without specifying the number of clusters in advance, the only need to do is setting the penalty parameter λ[7]. According to the increment of the penalty parameter, the number of clusters increases from 1 to k. Each clustering centers are the average of the sample points in it. The clustering results are more stable. The improved k-means clustering algorithm reduces the sensitivity to the initial clustering center, improves the convergence rate, and obtains better clustering results. The main steps of the improved k-means clustering algorithm are as follows:

Step 1. Initializing the number of clusters as k=1, and initializing the cluster center as the average of all samples.

Step 2. Calculating the distance from each sample point to the cluster center.

Step 3. When existing some points from which the distance to the center point is larger than the penalty parameter λ, let the number of clusters is k=k+1, and recalculating the distance from each sample point to k cluster centers. Those sample points will be assigned to the nearest class, and then the cluster centers are recalculated.

Step 4. Repeat the step 2 and the step 3 until the number k of clusters does not change, or the maximum number of iterations is reached. The algorithm ends.

3.2 The General Method to Determine the Penalty Parameters

The improved k-means clustering algorithm needs giving penalty parameter λ. Since the number k of clusters varies with λ, the value of λ can be determined by observing the change characteristics of k value. The corresponding value of λ while the trend of k value varies gently can be selected as the appropriate penalty parameter. The range of value λ is determined by the Euclidean distance from sample points to cluster centers, the minimum distance is the lower limit of λ, the maximum distance is the upper limit of λ.

3.3 Introduced the Degree of Compatibility of Decision Table to Determine Penalty Parameters

The k-means clustering algorithm is an unsupervised discretization method. When it is used to discretize a single sample attribute without considering the mutual relationship among attributes, even though the penalty parameter is introduced, the reduction result is not very effective. In order to improve the efficiency of reduction, in this paper, the degree of compatibility of the decision table is introduced in the clustering process, and the degree of certainty of decision table is used as the feedback information to maintain the degree of compatibility as a satisfactory value, so as to find the appropriate cluster center and partition the conditional attribute space. The inner relationship of the samples was well considered by the introduction of the degree of compatibility degree, and it is helpful to obtain the more reasonable partition. The main steps of algorithm are as follows:

Step 1. According to the selection rule of the penalty parameter λ,selecting the corresponding value λ when the number k of clusters is stable, the continuous attribute decision table is discretized initially;

Step 2. Calculating the degree of compatibility of the initial discrete decision table;

Step 3. Performing the following cyclic operations for each condition attribute;

Step 3.1. Changing the penalty parameters λ, the attribute is discretized again;

Step 3.2. Calculating the degree of certainty of this attribute to decision attribute;

Step 3.3. If the degree of certainty of the new discrete result is larger than the maximum of the certainty of the

attribute, the degree of compatibility of decision table will be calculated;

Step 3.4. If the degree of compatibility is larger than that of the initial discrete decision table, the decision table is updated.

The appropriate penalty parameters will eventual be obtained through the algorithm. The range of penalty parameters for each condition attribute can be chosen according to the trend graph of the number k of clusters to λ, that is，selecting the interval of λ values in which the number k tends to be stable.

4 Attribute Reduction Algorithm

The core theory of rough set is attribute reduction. In the discretization of continuous attribute data, introducing the degree of compatibility is to ensure the compatibility of decision table, and discretizing the continuous attribute decision table as a whole can improve the reduction efficiency of decision table. In this paper, an attribute reduction algorithm based on approximate decision entropy is adopted which redefines the attribute importance from the point of view of algebra and information to obtain smaller reduction result[10][11].

5 Experimental Test and Result Analysis

5.1 Example

The main steam temperature in thermal power plant was used as an example to verify the validity of the proposed algorithm mentioned above. The main steam temperature system is a complex thermal object, which has many influencing factors, and different variables have different degree of impact on it. In order to simplify the data structure and improve the performance, the factors which have little effect on the main steam temperature should be ignored.

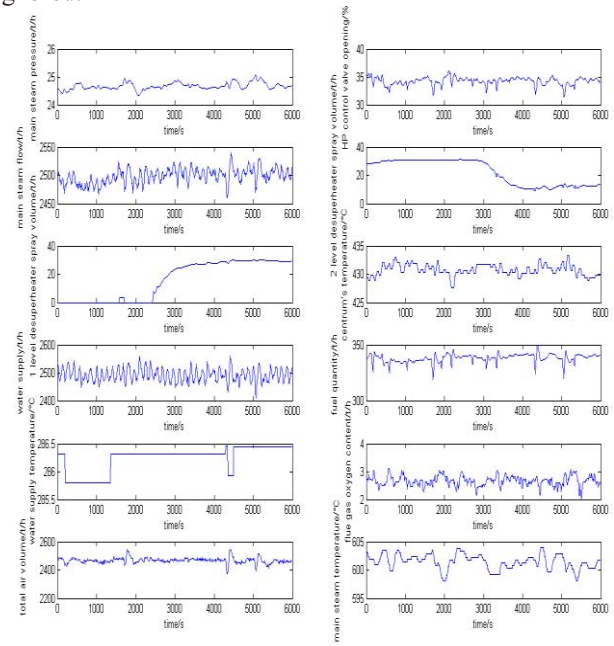

Fig.1:Main steam temperature system raw data curve

Since the sample space data of the main steam temperature system are all continuous attribute data, it is necessary to discretize the continuous attribute data before attribute reduction. According to the characteristic analysis, 11 variables that have an effect on the main steam temperature system were selected to constitute the condition attributes of the decision table, namely the main steam pressure, the main steam flow, HP control valve opening, 1st stage desuperheater spray volume, 2ed stage desuperheater spray volume, centrum's temperature, feedwater supply, fuel quantity, feedwater temperature, flue gas oxygen content and the total air volume[12]. The Main steam temperature is the decision attribute of decision table. The data curves of the main steam temperature system shown in Fig.1.

5.2 Data Preprocessing

The different attribute values in the sample data of the main steam temperature system adopted different dimensions. In order to eliminate the influence on the discretization results, firstly each attribute is normalized separately to a value between [0,1] according to the following expression:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where, x is the original attribute value, x^* is the normalized value of the attribute, x_{max} is the maximum value of the attribute, x_{min} is the minimum value of the attribute. Partial data after preprocessing were shown in Table 1. Among them, the condition attributes $C_1 \sim C_{11}$ represent the 11 variables affecting the main steam temperature, the decision attribute is the main steam temperature.

5.3 Determination of Penalty Parameters

Different penalty parameters will result in different clustering results, which will affect the efficiency of attribute reduction. Therefore, it is important to select reasonable penalty parameters for attribute reduction. In this paper, the relationship among the attributes of the actual system is fully considered. The degree of compatibility of the decision table is introduced as the judgment condition to decide whether the penalty parameter should be changed, which will provide the condition for the discretization of the continuous attribute data. The validity of the proposed method is verified by comparing it with the method which did not introduce the degree of compatibility.

The upper limit of the penalty parameter λ is the maximum of the Euclidean distance from the sample point to the initial cluster center, and the lower limit is the minimum of the Euclidean distance.

From Fig.2 we can see that when the number k of clusters tends to be stationary, selecting corresponding value as penalty parameters. The values of the penalty parameters of each condition attribute are shown in Table 2.

According to the trend graph in Fig.2, we can roughly judge the value of λ, which were shown in Table 2, but the choice is still subjective. And the discretization process is carried out on a single attribute, which does not take the correlation between attributes into account. But for the

whole system, each attribute is interrelated, the decision table after discretization should ensure the original degree of compatibility. Therefore, the degree of compatibility of the decision table was considered, which is used as the feedback information to select the optimal penalty parameter, which were shown in Table 3.

Table 1：Decision Table After the Data Normalization

No.	Condition attribute											Decision Attribute
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	
1	0.2653	0.6290	0.4836	0.8236	0	0.4873	0.2537	0.7425	0.5773	0.5232	0.3471	0.4998
2	0.2755	0.5968	0.4618	0.8236	0	0.4873	0.2467	0.7154	0.5773	0.5497	0.3866	0.4595
3	0.2755	0.5323	0.4341	0.7454	0	0.5372	0.3716	0.6367	0.5773	0.5232	0.3184	0.4595
4	0.2857	0.4032	0.3473	0.7454	0	0.5372	0.2746	0.5447	0.5773	0.5497	0.2802	0.4219
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
598	0.2653	0.4194	0.3204	0.1201	0	0.3193	0.4659	0.4231	0	0.5629	0.5019	0.5286
599	0.2857	0.3548	0.3315	0.1700	0	0.3193	0.3779	0.4713	0	0.5629	0.4598	0.5286
600	0.2653	0.4194	0.3778	0.3092	0	0.3193	0.3912	0.6044	0	0.6159	0.4349	0.5286

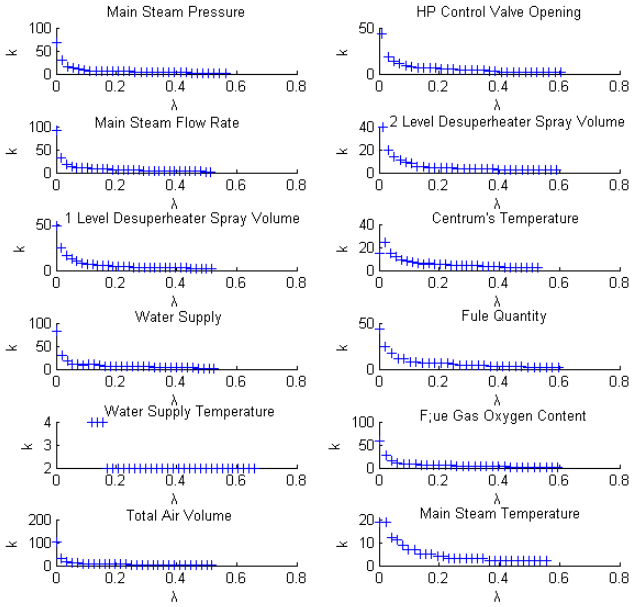

Fig.2:The trend graph of the number *k* varying with λ

Comparing the results of Table 2 and Table 3, we can see that the penalty parameter value obtained in the two cases is not exactly the same. In particular, some attributes are quite different, such as C_3、C_4、C_6、C_{11}.

5.4 Data Discretization

The values of Table 2 and Table 3 are taken separately as the given penalty parameters of the clustering algorithm to discretize the continuous attribute decision table. The resulting decision table under different λ were shown in Table 4 and Table 5.

Table2: The Value of λ According to the Trend Graph

Penalty Parameters		λ
Condition Attribute	C_1	0.4
	C_2	0.4
	C_3	0.1
	C_4	0.2
	C_5	0.2
	C_6	0.2
	C_7	0.2
	C_8	0.3
	C_9	0.3
	C_{10}	0.2
	C_{11}	0.2
Decision Attribute		0.4

Table 3:The Value of λ Considering the Compatibility

Penalty Parameters		λ
Condition Attribute	C_1	0.3840
	C_2	0.3910
	C_3	0.001
	C_4	0.0540
	C_5	0.12
	C_6	0.004
	C_7	0.3780
	C_8	0.488
	C_9	0.1610
	C_{10}	0.374
	C_{11}	0.001
Decision Attribute		0.4

Table 4: Decision Table After Discretization Without Considering the Degree of Compatibility

Universe	Condition Attribute											Decision
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	Attribute
1	1	1	3	2	2	2	7	1	1	1	1	1
2	2	1	2	2	2	2	4	1	1	1	1	1
3	2	1	3	2	2	2	4	1	1	1	1	1
4	2	1	2	2	2	2	4	1	1	1	1	1
5	2	1	2	2	2	2	1	1	1	1	1	1
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
596	1	1	1	1	1	2	1	3	1	3	2	1
597	1	1	3	1	1	2	1	3	1	1	2	1
598	1	1	3	1	1	2	1	3	1	1	1	1
599	1	1	2	1	1	2	2	3	1	1	2	1
600	1	1	5	1	1	2	4	3	1	3	2	1

Table 5: Decision Table After Discretization Considering the Degree of Compatibility

Universe	Condition Attribute											Decision
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	Attribute
1	1	1	22	4	2	6	2	1	1	1	29	1
2	1	1	28	4	2	6	2	1	1	1	22	1
3	1	1	27	4	2	6	2	1	1	1	24	1
4	1	1	48	4	2	6	2	1	1	1	43	1
5	1	1	14	4	2	6	1	1	1	1	49	1
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
596	1	1	23	10	1	11	1	1	1	3	26	1
597	1	1	20	10	1	11	1	1	1	1	26	1
598	1	1	37	10	1	11	1	1	1	1	61	1
599	1	1	31	10	1	11	2	1	1	1	31	1
600	1	1	6	10	1	11	2	1	1	1	31	1

5.5 Data Attribute Reduction and Result Analysis

By using the attribute reduction algorithm based on approximate decision entropy, the discretization decision tables, Table 4 and Table 5, are reduced and the redundant attributes are deleted. The result of comparison is shown in Table 6. It is visible that if the penalty parameters were set according to Table 2, reduced attributes were $[C_1,C_3,C_4,C_6,C_7,C_8,C_{10},C_{11}]$. The data structure is reduced from 11 dimension to 8 dimension, and if the penalty parameters were set according to Table 3, reduced attributes were $[C_3,C_6,C_{10},C_{11}]$. The data structure is reduced from 11 dimension to 4 dimension. By comparison, the penalty parameters obtained by introducing the degree of compatibility are more effective to discretize the continuous attributes, which improves the efficiency of the reduction and reduces the complexity of the data structure.

Table 6: Comparison of Reduction Results

Universe	The Attributes After Reduction Without Considering Compatibility	The Attributes After Reduction Considering Compatibility	Decision Attribute

	C_1	C_3	C_4	C_6	C_7	C_8	C_{10}	C_{11}	C_3	C_6	C_{10}	C_{11}	
1	1	22	4	6	2	1	1	29	22	6	1	29	1
2	1	28	4	6	2	1	1	22	28	6	1	22	1
3	1	27	4	6	2	1	1	24	27	6	1	24	1
4	1	48	4	6	2	1	1	43	48	6	1	43	1
5	1	14	4	6	1	1	1	49	14	6	1	49	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
596	1	23	10	11	1	1	3	26	23	11	3	26	1
597	1	20	10	11	1	1	1	26	20	11	1	26	1
598	1	37	10	11	1	1	1	61	37	11	1	61	1
599	1	31	10	11	2	1	1	31	31	11	1	31	1
600	1	6	10	11	2	1	1	31	6	11	1	31	1

6 Conclusion

In this paper, the compatibility of decision tables was introduced into the determination of penalty parameters in the improved k-means clustering algorithm. Taking the degree of compatibility as the feedback information of the clustering results ensured the compatibility of decision tables, which would be the compensation when this clustering algorithm was used to discretize those associated continuous attributes. And only the penalty parameter need giving in the algorithm. The algorithm was applied to discretize associated continuous attributes of the real thermal system. The results show that the combination of the degree of compatibility and the improved k-means clustering algorithm improves the efficiency of attribute reduction and greatly simplifies the data structure. The strategy can be used in data dimension reduction. It would be significant for system modeling which contains plentiful continuous attributes.

References

[1] Zhou Z G. Based on the reduction of attribute importance algorithm in the application of data mining research [J]. *Information Technology and Informatization*,2015,04: 199-200.

[2] Wan X, Hu N S, Han P F, et al. Research on application of big data mining technology in performance optimization of steam turbines[J]. *Proccedings of the CSEE*, 2016,36(02): 459-467.

[3] Zhang Y C, Su B H, Cao J. Study on application of attributive reduction based on rough sets in data mining[J]. *Computer Science,* 2013,40(08):223-226.

[4] Sun J, Hu M, Zhao J. An optimal algorithm for K-means initial clustering center selection[J]. *Journal of Changchun University of Technology*, 2016,37(01):25-29.

[5] Han Y M. Improved K-means dynamic clustering algorithm based on information entropy[J]. *Journal of Chongqing University of Posts and Telecommunication (Natural Science Edition)*, 2016,28(02):254-259.

[6] Zuo N N. K-means clustering method based on improved genetic algorithm[J]. *Software Guide,*2016,15(04):32-34.

[7] Zhou Q L, Lei J Y, Wang Y D, et al. A clustering algorithm with parameters that no need to determine the clustering number[J]. *Hebei Journal of Industrial Science and Technology*, 2015,32(02):123-128.

[8] Miao D Q, Li D G. R*ough set theory, algorithm and application*[M]. Beijing: Tsinghua University Press, 2008: 29-34

[9] Li S P, Chen X X, Xu J, et al. Generalization of the consistent decision table[J].*Journal of Gansu Sciences*, 2006,18(03):69- 71.

[10] Jiang F,Wang S S,Du J W,et al.Attribute reduct-ion based on approximation decision entropy[J].*Control and Decision*, 2015,30(01):65-70.

[11] Sun H R, Wang R, GENG J Y. The Thermal System Modeling Based on Entropy and BP Neural Network [J]. Journal of System Simulation,2017,(01):226-233.

[12] Hou X N. *Modeling and accuracy research of 1000MW USC Unit's main steam temperature*[D].North China Electric Power University,2015