A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data

Duo Chen, Du-Wu Cui, Chao-Xue Wang, Zhu-Rong Wang

School of Computer Science and Engineering, Xi'an University of Technology,
Xi'an, 710048, China
vkxtfj@163.net

Abstract

In this paper, rough set theory is applied to the clustering analysis. The clustering decision table is formed through the introduction of decision attribute into data table, thereby further defining the attribute membership matrix. The consistent degree and aggregate degree are present, and their functions in the clustering process are deeply analyzed. The clustering level calculation formula is designed, in which two factors such as consistent degree and aggregate degree are taken into comprehensive account. Also, this paper gives the categorical similarity measure based on Euclidean distance so as to better solve the problem of difficult measurement of categorical data because of the non-numerical data nature. On the basis of the above work, a novel categorical clustering algorithm is designed.

Keywords: Clustering; Categorical data; Rough set theory; Similarity measure.

1. Introduction

Clustering analysis is an important research project in knowledge discovery and data mining (KDDM). In practical application, the data sets contain numerical and categorical (nominal) data in general. Accordingly, clustering algorithm is required to able to deal with both numerical data and categorical data. K-means algorithm suggested by Mac Queen [1] is one of the most popular clustering algorithms, and it works only on numeric data. Accordingly, K-modes presented by Huang [2] has expanded K-means algorithm so as to deal with categorical data. The fuzzy K-modes algorithm put forward by Huang and Ng [3] has improved the clustering accuracy by using fuzzy processing technology. The above-mentioned algorithms belong to the partition methods in clustering analysis, with high operation efficiency, but there exist the shortcoming that clustering results are often dependent on the selection of the initial points. With an aim at the shortcoming, Bradley and Fayyad [4] posed the refining initial points for k-means clustering; Sun, Zhu and Chen [5] advanced the iterative initial-points refinement algorithm for categorical data clustering; and D. W. Kim, K. H. Lee and D. Lee [6] suggested the fuzzy clustering of categorical data using fuzzy centroids. Concept clustering algorithm [7,8,9] is another kind of method to deal with the clustering problems of categorical data such that this kind of method can not only realize the clustering process, but provide the concept descriptions of clusters as well. The hierarchical clustering algorithm can deal with both numerical data and categorical data. ROCK algorithm [10] is a type of agglomerative hierarchical clustering algorithm for categorical data. Rough set theory (RST) suggested by Pawlak [11,12] is a new mathematical tool to deal with vagueness and uncertainty, with successful applicable results obtained in many fields of information system. Particularly in recent years, being a kind of favorable mathematical tool, RST has displayed the vast applicable future in KDDM field. It can be expected that RST holds a vast applicable promise either in theoretical research field or in the practical application.

In dealing with the categorical data, one of the difficulties is to resolve the problem of similarity measure, for the nature of categorical data is the non-numerical so that Euclidean distance extensively-used in numerical data processing can not be employed directly. However, in this paper, RST is applied to clustering analysis; and the clustering data set is mapped as the decision table through introducing a decision attribute, whereby configuring the attribute membership matrix and presenting clustering consistent degree and aggregate degree and analyzing their functions in the clustering process. Also, the categorical similarity measure based on Euclidean distance is suggested. This measure is better to solve the problem of difficult measurement because of non-numerical nature of categorical data. Based on the analysis, this paper designs the rough set-based hierarchical clustering algorithm for categorical data. Theoretical analysis and experimental results indicate that this algorithm is valid.

2. Basic RST Notions

This section briefs on the basic notions of RST used in this paper and the detailed definitions can be referred to some related literatures [11,12,13].

Definition 2.1 An information system (IS, sometimes called data table, attribute-value system, etc.) is a pair (U,A), where U is a nonempty, finite set of objects called the universe and A is a nonempty, finite set of attributes, such that $a: U \rightarrow V_a$ for any $a \in A$, where V_a is called the domain of attribute a.

Each nonempty subset $B \subseteq A$ determines an indiscernibility relation

$$R_B = \{(x, y)\ U \times U: a(x)=a(y), \forall\ a \in B\ \}$$

R_B partitions U into equivalence classes

$$U/R_B=\{[x]_B: x \in U\}$$

where $[x]_B$ denotes the equivalence class determined by x with respect to (wrt) B, i.e.,

$$[x]_B = \{y \in U: (x, y) \in R_B\}$$

A decision table (DT) is an IS $(U, A \bigcup D)$, where $A \bigcap D=\emptyset$. Then term A is called the condition attribute set, and D is called the decision attribute set. If $R_A \subseteq R_D$, then $(U, A \bigcup D)$ is consistent, otherwise it is inconsistent. In general D has only one attribute d.

Definition 2.2 Let $(U, A \bigcup \{d\})$ be a DT, $B \subseteq A$, and $U/R_{\{d\}}=\{D_1,...,D_r\}$. A membership distribution function μ_B: $U \rightarrow [0,1]^r$ is defined as follows:

$$\mu_B(x) = \{\ D(D_1/[x]_B)\ , \dots ,\ D(D_r/[x]_B)\ \},\quad x \in U$$

where, $D(D_j/[x]_B\ ) =|D_j \bigcap [x]_B| \ /\ |[x]_B|$

3. Rough Set-Based Clustering

3.1 Related Definitions and Theorems

In this paper clustering problems are described using the clustering DT $(U,A\bigcup\{d\})$, in which U is the universe, one element $x_i\in U$ is called as an object, A is the attribute set, $A=\{a_1,\cdots,a_s\}$, and d is the introduced decision attribute with the domain $V_d\in\{1,...r\}$, $r\leqslant|U|$. The main purpose of introducing the decision attribute d is to use it to express the clusters. In terms of RST, the attribute d determines an indiscernibility relation $R_{\{d\}}$, which partitions U into equivalence classes, $U/R_{\{d\}}=\{D_1,...,D_r\}=P$, named as clustering model in this paper. It can be considered that the clustering model P expresses a kind of clustering result.

Definition 3.1. Let $(U, A\bigcup\{d\})$ be a DT, and $P=U/R_{\{d\}}=\{D_1,\ldots,D_r\}$, $A=\{a_1,..., a_s\}$, $n=|U|$, an attribute membership matrix M_k is defined as follows:

$M_k=[\mu_k(i,j)]=\{D(D_j/[x_i]a_k)\}$, $i\in\{1,...,n\}$, $j\in\{1,...,r\}$, $k\in\{1,...,s\}$

Obviously, $M_k=[\mu_{\mathrm{k}}(i,j)]$ is $n\times r$ matrix and $\mu_{\mathrm{k}}(i,j)\in[0,1]$. Definition 3.1 comes from Definition 2.2, which can be considered as a special example when the condition attribute subset $B\subseteq A$ in Definition 2.2 takes the individual condition attribute $a_k\in A$. It is just because the individual conditional attribute is taken that the membership matrix M_k has the different features from those of membership distribution function as well as different manipulating methods.

Theorem 3.1 Let $(U, A\bigcup\{d\})$ be a DT, and $A=\{a_1,...,a_s\}$, $n=|U|$, if $P=U/R_d=\{D_1,...,D_n\}=\{\{x_1\},...,\{x_n\}\}$, and then,

(1) M_k is n-order symmetric matrix, i.e., $\mu_{\mathrm{k}}(i,j)=\mu_k(j,i)$, $\forall k\in\{1,...,s\}$

(2) $\sum_{j=1}^{r}\mu_k(i,j)=1$, *and* , $\sum_{i=1}^{n}\mu_k(i,j)=1$, $\forall\ k\in\{1,...,s\}$

Proof: It follows immediately from Definition 3.1. ■

Theorem3.2 (Mergence Theorem) Let $(U, A\bigcup\{d\})$ be a DT, and $A=\{a_1,...,a_s\}$, $P=U/R_{\{d\}}=\{D_1,...,D_r\}$, $n=|U|$. If $\forall f,\ g\in\{1,...r\}$, $f\neq g$ and $D'=D_f\bigcup D_g$, and then $D(D'/[x_i]a_k)=D(D_f/[x_i]a_k)+D(D_g/[x_i]a_k)$, $\forall i\in\{1,...,n\}$, $k\in\{1,...,s\}$

Proof. $\forall k\in\{1,...,s\}$ and $i\in\{1,...,n\}$, we have:

$D(D_f/[x_i]a_k)+D(D_g/[x_i]a_k)=|D_f\bigcap[x_i]a_k|/|[x_i]a_k|+|D_g\bigcap[x_i]a_k|/|[x_i]a_k|$

$=[\,|D_f[x_i]a_k|+|D_g[x_i]a_k|\,]/|[x_i]a_k|=|D'[x_i]a_k|/|[x_i]a_k|$ *(since* $D_f\bigcap D_g=\emptyset$ *)*

$=D(D'/[x_i]a_k)$ ■

In terms of Mergence Theorem, the merging operation in clustering process can be achieved via the corresponding addition operation in attribute membership matrix, being suitable to either the cluster or the individual object. Therefore, Mergence Theorem is of great importance for clustering algorithm.

Deduction 3.2 Let $(U,A\bigcup\{d\})$ be a DT, and $P=U/R_{\{d\}}=\{D_1,...,D_r\}$, $A=\{a_1,...,a_s\}$, $n=|U|$. If $f,\ g\in\{1,...,r\}$, $f\neq g$ and $D'=D_f\bigcup D_g$, then $\forall i\in\{1,...,n\}$, $k\in\{1,...,s\}$, we have

(1) $D(D'/[x_i]a_k) \geqslant D(D_f/[x_i]a_k)$ *and* $D(D'/[x_i]a_k) \geqslant D(D_g/[x_i]a_k)$,

(2) $\sum_{i=1}^{n} \mu_k(i,k')=|D'|$, where k' is the sequence number of the cluster D'.

3.2 Consistent Degree and Aggregate Degree

In the case of traditional RST, the consistent DT and inconsistent DT are only given, and the consistent measure of the DT is not defined. In fact, in a clustering DT ($U, A \bigcup \{d\}$), for each $a_k \in A$, the *jth* column of M_k represents the membership distribution of each object in U corresponding to the cluster D_j in clustering model P. This paper holds that this distribution can include the information of coordination of the condition attribute set A to the clustering model P in the DT, further reflecting that the clustering accuracy can be used as a kind of measure in clustering process. On the basis of the above analysis, we have offered the consistent degree definition.

Definition 3.2 Named mapping $\Phi:[0,1] \rightarrow [0,1]$ as the coordination, if satisfying the following properties:

(1) $\Phi(x) = 1$ *iff* $x = 0, 1,$ and $\Phi(1/2)=0$

(2) $\Phi(x) = \Phi(1-x)$

(3) $\Phi(x)$ on $[0, 1/2]$ monotonic reduction

In Definition 3.2, condition (1) indicates that when independent variable x is 1 or 0, i.e. the object completely or incompletely falls under the specified cluster, the coordination can reach its maximum value 1; correspondingly, when x is $1/2$, i.e. the membership relation is entirely unclear such that the coordination reaches its minimum value 0; condition (2) indicates that the coordination is about $x=0.5$ symmetry; condition (3) specifies monotonic features of the mapping.

We introduce $\Phi(x) = 2 \times |x - 0.5|$ served as the coordination function.

Definition 3.3 Let $(U,A \bigcup \{d\})$ be a DT, $P=U/R_{\{d\}}=\{D_1,..,D_r\}$. We define the consistent degree of the D_j in P as:

$$CON(P,j)=\frac{1}{s \times n}\sum_{k=1}^{s}\sum_{i=1}^{n}\Phi\,[\,\mu_k(i,j)] \qquad \textbf{(1)}$$

and the consistent degree of the DT as:

$$CON(P)=\frac{1}{r}\sum_{j=1}^{r}[\ CON(P,j)\ ] \qquad \textbf{(2)}$$

Obviously, $CON(P,j),\ CON(P) \in [0,1]$

Theorem 3.3 In $(U,A \bigcup \{d\})$, if $P=U/R_{\{d\}}=\{D_1,...,D_r\}=\{U\}$, then $CON(P)=1$

Proof. It follows immediately form the definition of $CON(P)$ ■

Without considering the case of $P= \{U\}$, the larger the $CON(P)$ is, the higher the clustering accuracy is. The clustering process should render $CON(P)$ to advance in the direction of enlargement. But it is not complete to guide the clustering via $CON(P)$, this is because consistent degree does not include the clusters in P upon the objects containing information. This paper ushers the aggregate degree to indicate that the clusters contain the objects to a certain extent.

Definition 3.4 Let $(U,A\cup\{d\})$ be a DT, $n=|U|$, $A=\{a_1...,a_s\}$, $P=U/R_{\{d\}}=\{D_1,...,D_r\}$, we define the aggregate degree of D_j in P as:

$$AGD(P,j)=\frac{1}{s}\sum_{k=1}^{s}\Psi\{\ [\ \frac{1}{n}\sum_{i=1}^{n}\mu_{\mathrm{k}}(\mathrm{i,j})^{2}\ ]^{1/2}\ \} \qquad \textbf{(3)}$$

where, $\Psi(\cdot)=[1-ln(\cdot)]^{-1}$, and the aggregate degree of the DT as:

$$AGD(P)=\frac{1}{r}\sum_{j=1}^{r}[\ AGD(P,j)\ ] \qquad \textbf{(4)}$$

In Definition 3.4, the mean-root-square is employed to define $AGD(P,j)$, whereas the arithmetic mean value is not adopted to do the definition. This is mainly due to avoiding yielding trivial values (i.e. simple 0 or 1). This method borrows the predictiveness definition method listed in literature [9]. But the predictiveness and aggregate degree have three points in difference: the first is that the predictiveness is defined in terms of probability distribution, while the aggregate degree is defined in terms of membership matrix; the second is that the aggregate degree has had the normalization manipulation, whereby avoiding that the aggregate degree can quickly reach rather large numerical values with an increase in cluster numbers; and the third is that ushering in Ψ function is convenient to match with consistent degree. As far as Ψ function is concerned, this section will analyze it in details in later part.

Theorem 3.4 Let $(U,A\cup\{d\})$ be a DT, $P=U/R_{\{d\}}=\{D_1,...,D_r\}$, then $AGD(P)\in(0,1]$

Proof. For any $a_k\in A$, since $\mu_{\mathrm{k}}(i,j)\in[0,1]$, we have $AGD(P)\leqslant 1$; Because if elements in any column of M_k are all zero, indicating that any objects in U do not fall under the cluster corresponding to this column in P, this is impossible, so we have $AGD(P)>0$. To sum up the above descriptions, $AGD(P)\in(0,1]$ holds. ■

Theorem 3.5 Let $(U,A\cup\{d\})$ be a DT, $P=U/R_{\{d\}}=\{D_1,...,D_r\},$. and $D'=D_f\cup D_h$, $\forall f,h\in\{1,...r\}$,$f\neq h$. If $P'=(P-\{D_f, D_h\})\cup\{D'\}$, then $AGD(P')>AGD(P)$

Proof. It follows immediately form Definition 3.4 and Merge Theorem. ■

In terms of Definition 3.4 and Theorem 3.5, $AGD(P)$ expresses the containing degree of clusters in P upon the objects. In agglomerative hierarchical clustering algorithm, from the starting P including n clusters, i.e. each object being a cluster to the final P containing one cluster, i.e. all the objects being a cluster, has formed the dynamic clustering map. In the dynamic clustering process, aggregate degree goes up monotonically but the consistent degree maybe not. Actually in practical application, the consistent degree has some reductions with an increase in aggregate degree. The nature of monotonic increase of aggregate degree is of important application value. We may appoint the threshold $\lambda\in(0,1)$ of the aggregate degree as the algorithm ending condition. An effective algorithm should take the two indexes of $CON(P)$ and $AGD(P)$ into comprehensive consideration, whereby rendering $CON(P)$ to have some increase or to have minimal decrease with an increase in $AGD(P)$ in the clustering process, for this reason the calculation formula of clustering level is introduced.

Definition 3.5 Let $(U, A\bigcup\{d\})$ be a DT, $P=U/R_{\{d\}},=\{D_1,...D_r\}$, we define the clustering level of D_j in P as:

$$LEV(P,j) = [2\times CON(P,j)\times AGD(P,j)] / [CON(P,j) + AGD(P,j)] \qquad \textbf{(5)}$$

and the clustering level of the DT as:

$$LEV(P) = [2\times CON(P)\times AGD(P)] / [CON(P) + AGD(P)] \qquad \textbf{(6)}$$

Clustering level can comprehensively reflect two indexes of the consistent degree and the aggregate degree. It is easy to proof that $LEV(P)$ is the monotonic increasing function wrt $AGD(P)$ and $CON(P)$. Obviously, $LEV(P)\in[0,1]$, and $LEV(P)=0$, iff $CON(P)=0$; $LEV(P)=1$, iff $CON(P)=AGD(P)=1$. It can be known from Definition 3.5 that $AGD(P)$ and $CON(P)$ in $LEV(P)$ are two symmetrical parameters. When $AGD(P)$ and $CON(P)$ values are closer, they will have the similar effect upon $LEV(P)$; when the value of one of the parameters is much smaller than that of other parameter, the parameter having the small value will have the greater effect upon $LEV(P)$.

The function Ψ is introduced into $AGD(P)$ definition (Definition 3.4). This is because in the initial stage of agglomerative algorithm, $AGD(P)$ values are small (or very small when the data set are very large) so that the main purpose of introducing function Ψ lies in raising $AGD(P)$ value in the initial stage of algorithm so as to match with $CON(P)$. Without introducing function Ψ and in the initial stage of algorithm, $AGD(P)$ in the clustering level is too small to render $CON(P)$ to play its role, which is no doubt to affect the accuracy of clustering algorithm.

3.3 The Categorical Similarity Measure Based on Euclidean Distance

In this section, a novel categorical similarity measure is suggested.

Definition 3.6 In $(U,A\bigcup\{d\})$, $A=\{a_1,...,a_s\}$, $P=U/R_{\{d\}}=\{D_1,...,D_r\}$, $n=|U|$, $\forall f,h\in\{1,...,r\}$, the similarity between two clusters D_f and D_h in P can be defined as follows:

$$SIM(f,h)=\frac{1}{s}\sum_{k=1}^{s}\{1-[\frac{1}{n}\sum_{i=1}^{n}(\mu_k(i,f)/|D_f|-\mu_k(i,h)/|D_h|)^2]^{1/2}\} \qquad (7)$$

Definition 3.6 cites Euclidean distance to describe the similarity among clusters such that the greater the value of distance is, the smaller the value of similarity is. Not only can the similarity defined using Euclidean distance measure the numbers of the same attributes between two clusters and differences in the numbers of dissimilar attributes, but also the key lies in expressing the degrees of the similar and dissimilar attributes, whose nature is to do numerical processing of categorical attributes. In *Eq.(7)*, $\mu_k(i,f)/|D_f|$ *and* $\mu_k(i,h)/|D_h|$ indicate the centers of cluster D_f and cluster D_h respectively, i.e. using membership mean value represents the cluster centers in such a way as to borrow the expressing method of the well-known K-means algorithm [1] in numerical clustering. If D_f and D_h contain one object, then $|D_f| = |D_h|=1$. Accordingly Definition *3.6* is also adaptable to this situation so that Definition 3.6 expresses the similarities among the cluster versus the cluster, and the object versus the cluster as well as the object versus the object.

3.4 The Algorithm

This section presents the Rough Set-Based Agglomeration Hierarchy Clustering Algorithm (RAHCA). This algorithm first lets $P=U/R_{\{d\}}=\{\{x_1\},...,\{x_n\}\}$, and then conducts merging operation in terms of the clustering level and similarity measure. This operation can be carried out till the number of clusters m or the aggregate degree threshold λ given by users and outputs the clustering results, i.e. clustering model P. This algorithm can also be conducted till all the objects merged into one cluster. In such a way the algorithm outputs the dynamic clustering map.

```
Algorithm RAHCA
Input: Data table $(U,A)$, number of clusters $m$ (or aggregate degree threshold $\lambda$)
Output: Clustering model $P$
Step 1 Let $P(0)= \{D_1,...,D_r\}=\{\{x_1\},...\{x_n\}\}$, $n=|U|$, $r=n$.
Step 2. In terms of Definition 3.1, derive the membership matrix $M_k$, $k\in\{1,...,|A|\}$
Step 3. Repeat the following operations, until $r = m$ (or $LEV(P) \geqslant\lambda$)
  Step 3.1 Find the cluster $D_{min}$ with the minimum $LEV$ value using $Eq.(5)$.
  Step 3.2 Compute the similarity between $D_{min}$ and the rest of the clusters in $P$
           using $Eq.(7)$, let $D_{sim}$ be the cluster scoring the highest similarity with $D_{min}$.
  Step 3.3 Update clustering model $P$ and membership matrix $M_k$, $k\in\{1,...,|A|\}$ by
           merging $D_{min}$ and $D_{sim}$ according to Mergence Theorem, and then r =r-1.
```

The algorithm complexity is analyzed as follows. In step 2, the equivalence classes $U/R_{\{a_k\}}$ should be first computed for computation M_k, and then $r\times|U|$ elements in M_k must be conducted, where r is the number of the clusters in P with the maximum $|U|$. Algorithm *1* in literature[14] can be used to compute equivalence classes, whose time complexity is $O(|A||U|)log(|U|)$, thereby the time complexity of step 2 should be $O(|A||U|log|U|)+O(r|A||U|)$. In step 3, the maximal number of the iterations is $|U|-1$, and each iteration needs to compute LEV r times in step 3.1 and SIM $r-1$ times in step 3.2, both of them can be computed in $O(r|A||U|)$, and the updating operations in step 3.3 can be finished in $O(|A||U|)$, such that the time complexity of step 3 is $O(r|A||U|^2)$. So the time complexity of the algorithm can be estimated as $O(|A||U|^3)$. We can easily analyze that the space complexity is $O(|A||U|^2)$. It is worth mentioning that the above-mentioned analysis is in the most unfavorable situation and regards M_k as the dense matrix, but in practical application M_k is the highly sparse matrix. Using a type of appropriate sparse matrix operating strategy, both the time complexity and the space complexity of the algorithm will be reduced notably.

4. Experimental Results

4.1 Artificial Data Set

In this section, one simple example is used to demonstrate the execution of RAHCA. In the clustering DT shown in Table 1, the universe $U=\{x_1,...x_5\}$, the initial categorical condition attribute set $A=\{a_1, a_2, a_3\}$, d is the introduced decision attribute with the domain $V_d=\{1,...,5\}$, and it is in its initial state as shown in Table 1.

Table 1. A Clustering DT

U	a_1	a_2	a_3	d
x_1	1	1	1	1
x_2	2	2	3	2
x_3	1	1	2	3
x_4	2	2	1	4
x_5	3	2	3	5

In the first iteration, the *LEV* values of D_j, $j\in\{1,...,5\}$ are *0.523, 0.515, 0.591, 0.515* and *0.581* respectively, $LEV[P(0),2]=0.52$ is the minimum, let $D_{min}=D_2$, the *SIM* values between D_2 and the rest objects are *0.57, 0.53, 0.85* and *0.82* respectively, obviously the nearest cluster is D_4, $SIM(2,4)=0.85$. By merging D_2 and D_4, $P(1)=\{x_1,\{x_2,x_4\},x_3,x_5\}$ can be obtained. In terms of the same computing, in the second iteration by merging D_1 and D_3, $P(2)=\{\{x_1,x_3\},\{x_2,x_4\},x_5\}$ can be obtained. In the third iteration, $P(3)=\{\{x_1,x_3\},\{x_2,x_4,x_5\}\}$; and in the last iteration, $P(4)=\{\{x_1,x_2,x_3,x_4,x_5\}\}$, thus the algorithm comes to its end. It is easy to analyze that *P(2)* and *P(3)* are the rational results.

The dynamic clustering map is shown in Fig.1, and the curves that indicate the changing in numerical values of *CON(P)*, *AGD(P)* and *LEV(P)* in each iteration are shown in Fig. 2.

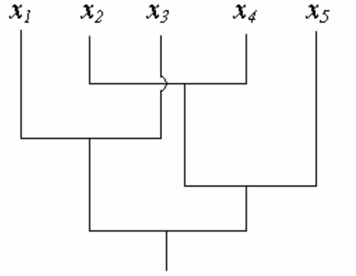

Fig. 1. Dynamic clustering map

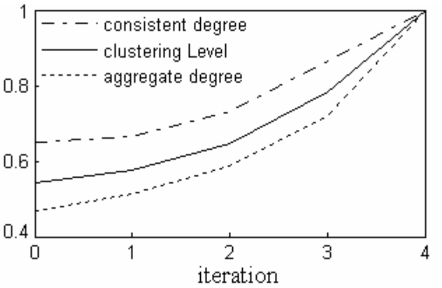

Fig. 2. Values of consistent degree, aggregate degree and clustering level in each iteration

4.2 UCI Data Sets

We ran experiments on 3 data sets obtained from the UCI Machine Learning Repository[1]. The Balloon date set contains 20 instances and each instance has 4 categorical attributes. It is classified into 2 classes. The Soybean date set contains 47 instances on diseases in soybeans and each instance has 35 categorical attributes. The data set is classified into 4 classes according to its disease type. The Voting data set derives from 1984 United States Congressional Voting Records Database. It contains 435 instances, which represent the voting records of 267 democrats and 168

[1] http://www.ics.uci.edu/~mlearn/MLRepository

republicans respectively. Each instance has 26 Boolean attributes. There are some missing attribute values in the date set and they are regarded as a special constant value in this paper.

The algorithm in literatures [2,9] and RAHCA algorithm proposed in this paper are adopted respectively (algorithm A, algorithm B and algorithm C for their short forms). The experiment results are shown in Table 2. In experiments the standard numbers of clusters are specified. The accuracy of an algorithm is the ratio of the total number of instances occurring in both the *ith* cluster and its corresponding true class to the number of instances in the data set. The results of Algorithm *A* are the average accuracy value of 50 runs. The experiment results indicate that the Algorithm *C*, i.e. RACHCA, has the highest accuracy for all of above date sets.

Table 2. Experiment results

Data set	Number of objects	Number of condition attributes	Number of clusters	Accuracy of algorithm *A* (%)	Accuracy of algorithm *B* (%)	Accuracy of algorithm *C* (%)
1.Balloon	20	4	2	69.8	80.0	100
2.Soybean	47	35	4	79.4	100.0	100
3.Voting	435	16	2	85.6	87.1	88.3

5. Conclusions

This paper applies RST to the clustering analysis in KDDM, and introduces the decision attribute to configure the clustering DT, whereby defining the membership matrix. This paper suggests the consistent degree and aggregate degree measures corresponding to the clustering model *P* of the DT. The two measures express two aspects of clustering process respectively. The consistent degree expresses the coordination degrees among the equivalence classes of the condition attribute set and the clusters in clustering model *P* of the DT, whereas aggregate degree indicates the containing degree of cluster itself upon the objects. In order to take the effect of the two factors upon the clustering process into comprehensive consideration, the clustering level calculation formula is designed, in which the consistent degree and aggregate degree are in the symmetric positions, whereby rendering the clustering process to be able to give consideration to the two measures at the same time. When their values are near, they can affect the clustering level in common; when one of which is smaller, this measure will play an important role in the clustering level. It is just for this reason that it can become a major factor to affect the clustering direction. In practical application, the aggregate degree is small in the initial clustering stage, so the aggregate degree is a major influencing factor in algorithm, and it is just at this time that the aggregate degree in algorithm should be raised. With the clustering ongoing, the aggregate degree increases gradually, and it, together with the consistent degree, guides the clustering. By the late stage of clustering, the consistent degree becomes a major influencing factor because of its becoming smaller such that it is just at this time that algorithm should first guarantee the consistent degree.

Also, this paper poses the definition of similarity measure for categorical data based on Euclidean distance, whereby the problem of difficult comparison of similarity caused by the nature of non-numerical values in categorical data can be better resolved. This measure is adaptable to the comparison among the cluster versus the cluster and the cluster versus the object as well as the object versus object. In addition, the consistent degree, aggregate degree and similarity measure are normalized in such a manner as to be convenient for the manipulation of each parameter and to make these parameters more coordinated and rationally brought into full play.

Based on the above analytical results, this paper designs the Rough set-based Agglomerative Hierarchical Clustering Algorithm (RAHCA) suitable for categorical data. This algorithm needs the users to offer the number of clusters or the aggregate degree threshold. Without these parameters specified, the algorithm outputs the dynamic clustering map which needs to be further analyzed so as to obtain the clustering results. The future research work will include the further improvement of algorithm efficiency as well as research on clustering algorithm for mixed numeric and categorical data.

References

[1] Mac Queen, J, "Some methods for classification and analysis of multivariate observations", Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability, 1967, pp. 281-297

[2] Huang, Z., "Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery Ⅱ. 1998, pp. 283-304

[3] Huang, Z., Ng, M.K., "A Fuzzy K-Modes Algorithm for Clustering Categorical Data", IEEE Trans. Fuzzy Systems, vol. 7, 1999, pp. 446-452

[4] Bradley, P., Fayyad, U., "Refining initial points for k-means clustering", Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, Los Altos, CA, 1998

[5] Sun,Y., Zhu, Q.M., Chen, Zh. X., "An iterative initial-points refinement algorithm for categorical data clustering", Pattern Recognition Letters, vol. 23, 2002, pp. 875-884

[6] Kim, D.W., Lee, K. H., Lee. D., "Fuzzy clustering of categorical data using fuzzy centroids". Pattern Recognition Letters, vol. 25, 2004, pp. 1263-1271

[7] Michalski R., Stepp, R., "Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy", IEEE Trans. Pattern Anal. Machine Intell., vol. 5, 1983, pp. 369-410

[8] Fisher, D., "Knowledge Acquisition via Incremental Conceptual Clustering". Machine Learning, vol. vol. 2, 1987, pp. 139-172

[9] Luis Talavera and Javier Béjar, "Generality-Based Conceptual Clustering with Probabilistic Concepts", IEEE Trans. Pattern Anal. Machine Intell., vol. 23, 2001, pp. 196-203

[10] Guha, S., Rastogi, R., Shim, K., Rock: "A Robust Clustering Algorithm for Categorical Attributes", Proc. Int. Conf. Data Engineering (ICDE'99), Sydney, Australia, 1999, pp. 512-521

[11] Pawlak, Z., "Rough sets", International Journal of Computer and Information Science, vol. 11, 1982, pp. 341-356

[12] Pawlak, Z,: "Rough Sets: Theoretical aspects of reasoning about data", Kluwer Academic Publishers, London , 1991

[13] Zhang, W.X., Mi, J. Sh., Wu, W. Zh., "Approaches to Knowledge Reductions in Inconsistent Systems", International Journal of Intelligent System, vol. 18, 2003, pp. 989-1000

[14] Liu Sh. H. et al: "Research on efficient algorithms for rough set methods", Chinese Journal of Computers, Vol. 26, 2003.5, pp.524-529(in Chinese)

Duo Chen, received his B.Eng’s degree in 1984 from Hebei University of Technology, Tianjin, China , received the M.eng’s degree in 1995 from Yanshan University, Qinhuangdao, China. He is an associate professor in the Computer Center of Tangshan College, Tangshan, China. He is currently working toward the Ph.D degree in computer science and engineering in Xi’an University of Technology, Xi'an, China. His research interests include data mining, intelligent computing and power electronics and electric drive, etc.

Du-Wu Cui, received his B.Eng’s degree and M.eng’s degree from Xi’an Jiaotong University, Xi'an, China in 1967 and 1983 respectively. He has been professor and Ph.D. supervisor in Xi’ an University of technology, Xi'an, China. His research interests include AI, multimedia technology and Decision Support System (DSS), etc.