Proceedings of the 27th Chinese Control Conference
July 16-18, 2008, Kunming,Yunnan, China

A Hybrid Clustering Algorithm Based on Grid Density and Rough Sets

Lv Huigang，Teng Peng，Huang Jun, Zhang Fengming

Inst. of Engineering, Air Force Engineering Univ., Xi'an 710038 , P. R. China
E-mail:lvhgzz@163.com

Abstract: According to the characters of dynamic and SOM clustering algorithm, propose a novel clustering method, rough dynamic clustering based on grid-density algorithm (GDRDC). The algorithm contains initial clustering stages and precise adjustment stages. During switch from the first stage to second stage, according to rough sets idea, class kernel and freedom point sets base on grid-density are determined, and though which the two stages are joined. Then making farther adjustment by dynamic clustering method, the final clustering result is get. The experiment result shows that it is better than SOM and K-means, especially for nonlinear separable data.

Key Word: Clustering Analysis, Grid-Density, SOM, Dynamic Clustering, Rough Sets

1 INTRODUCTION

Clustering analysis technology is a unsupervised learning method and an important research field in machine learning and data mining[1]. The main purpose is finding structure character and underlying classification information in sample data sets. According to rough-sets theory, the classification information is the knowledge on research object, it can be used to forecast and classify new data. Researchers have presented variety of clustering algorithms[2,3], the main classification on cluster algorithms are hierarchical clustering, partitioning clustering, density clustering and grid clustering. In a certain extent, these algorithms can accelerate clustering velocity and improve efficiency by means of feature selection[4], decreasing the number of dimension[5], or using reference points[6] , etc. But there still exist some contradictions and problems in cluster analysis by these algorithms, such as contradiction between cluster precision and efficiency, standard of initial value setting and so on. To solve these problems, in recent years, some algorithms of artificial intelligence are introduced to cluster analysis, such as SOM(Self Organizing Map), evolutionary algorithms, etc.

In this paper, we proposed a multidimensional data rough dynamic clustering algorithm based on grid-density (GDRDC). The process of our clustering algorithm can be partitioned into three stages. The objective of the first stage is compartmentalizing grid of multidimensional data space, and defining grid density. The second stage is clustering sample data by SOM and gain initial clustering result, determine class kernel of all initial classes according to rough idea. The last stage is classifying free sample points out of all class kernels to relevant initial class by dynamic clustering algorithm. The algorithm has high precision and strongly clustering capability on not linearly separable data sets, which is validated by an experiment on Iris data sets.

2 RELATED RESEARCH

2.1 Clustering Algorithm Based On Density and Grid

Density based Clustering algorithm and grid based clustering algorithm embody two different data description idea. Density based Clustering algorithm emphasize the data density distribution in sample space. It confirms classification though dividing high density field by low density, so it can be classify to partitioning clustering. The representational algorithms have DBSCAN algorithm, OPTICS algorithm and DENCLUE algorithm. Grid based clustering algorithm store dataset in grid data structure. According to grid structure character, the algorithm make grid unit as handle objects to implement clustering operation. The representational algorithms have STING algorithm, BANG algorithm and MAFIA algorithm. Some researcher also proposes clustering algorithm based on density and grid, such as CLIQUE algorithm and SUDBC algorithm.

DBSCAN algorithm group objects to clusters based on the density of data. It can use to mine clusters having arbitrary shapes and it classifies objects according to the data density in their neighborhoods. If the neighborhood of radius ε of some object contain more than k objects, then the object and its neighbors form a cluster. The cluster speed of DBSCAN algorithm is faster than hierarchical clustering. But the important issue of DBSCAN is its computational speed and efficiency with the large data sets, it is sensitive to ε .

CLIQUE algorithm is a cluster algorithm based on grid and density. The algorithm regards a cluster as a region that has a higher density of points than its surrounding region. To approximate the density of the data points, partitioning each dimension into the same number of equal length intervals and get many units with the same volume. Therefore, the number of points inside it can be regard as the density of the unit. The data points are separated according to the valleys of the density function. The clusters are unions of connected high density units within a subspace. CLIQUE automatically finds

subspaces with high-density clusters and produces identical results irrespective of the order in which the input records are presented, and it does not presume any canonical distribution for the input data[8]. CLIQUE is a high efficiency algorithm. It can handle multidimensional and noise data. But it rarely considers data distribution, which affect cluster quality in a certain extent.

2.2 Self-organizing Map Neural Network

The self-organizing map(SOM) is a self-organizing neural network algorithm based on unsupervised competitive learning. SOM proposed by T. Kohonen in 1981. The beneficial features of SOM are simpleness, availability and easy to visualize, which make it a useful method in data mining and KDD, So, It has been successfully applied in various engineering applications in pattern recognition, image analysis, process monitoring and fault diagnosis. It implements an ordered dimensionality reducing mapping of the training data. The map follows the probability density function of the data and is robust to missing data.

The SOM consists of M neurons located on a regular low-dimensional grid, usually one or two dimensional. Higher dimensional grids are possible, but they are not generally used since their visualization is problematic. The lattice of the grid can be either hexagonal or rectangular. Typical topologic structure like as Fig.1

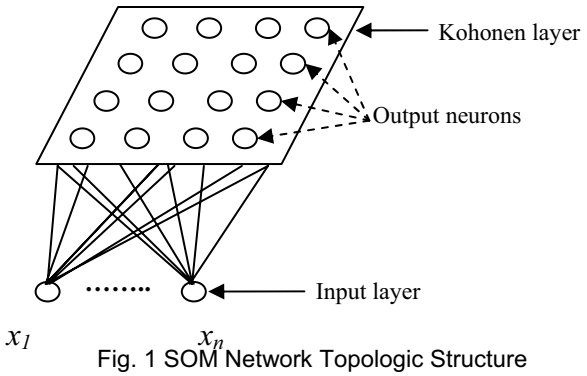

Fig. 1 SOM Network Topologic Structure

SOM consists of input layer and output layer, all neurons between two layers connect each other, if input vector x number is q, $x_i = [x_{i1}, x_{i2}, \ldots, x_{i3}] \in \mathbf{R}_n$, where n is the dimension of the input space, then connect weight vector can be defined as $w_j = [w_{j1}, w_{j2}, \cdots, w_{jn}] \in \mathbf{R}^n$. The basic SOM algorithm is iterative, at each training step t, a sample data vector $x(t)$ is randomly chosen from the training set. For each object i in the training data set, distance between $x(t)$ and all weight vectors w_i are computed. The winning neuron is the neuron with the weight vector closest to $x(t)$, which is denoted by c:

$$c = \arg \min_j \left\| x(t) - w_j \right\|$$

Initial weight w is set randomly, so it need update. We define $h_{jk}(t)$ as the neighborhood function around the winning neuron c at time t, neighboring nodes set is denoted as Nc. When neighborhood is rectangular, $h_{jk}(t)$ usually can be taken as a Gaussian function:

$$h_{jk}(t) = \exp\left\{ -\frac{\left\| r_j - r_c \right\|^2}{2\sigma^2(t)} \right\}$$

Where r_j and r_c denote as respectively position of neuron j and winning neuron c at time t. Width coefficient is denoted as $\sigma(t)$, $\sigma(t)$ decreasing when t increasing. Then, the weight update rule in SOM algorithm can be written as

$$w_j(t+1) = \begin{cases} w_j(t) + \eta(t) \cdot h_{jk}(t) \cdot (x(t) - w_j(t)), & j \in N_c \\ w_j(t), & otherwise \end{cases}$$

Where $\eta(t)$ is learning velocity at time t. SOM algorithm is can map discretionary dimension input model to one or two dimension discrete graphics in theory, moreover, it can preserve its essential topologic structure. The main step for training the SOM can be summarized as follows:

(1) For each input vector i, find out corresponding winning neuron c from among the nodes of the map, $c = \arg_j \min \left\| x(t) - w_j \right\|$.

(2) Find out neighborhood range Nc of winning neuron c, then adjusting weight vector of neuron among Nc, the process make weight vector of neuron among Nc more and more close with input vector i.

Learning velocity η and neighborhood range Nc decrease gradually along with iterative time, all neurons among kohonen layer separate each other and represent respectively one model of input space, which is the basic self-organizing cluster theory of SOM. Since the algorithm is proposed, many improved methods are researched, there are many variants of SOM, such as Growing SOM, TS-SOM and GA-SOM[7]. In this paper, SOM is used to initial cluster process, we look it as one rough classification knowledge, so the basic SOM algorithm is meet.

3 ROUGH DYNAMIC CLUSTERING ALGORITHM BASED ON GRID-DENSITY

3.1 Conception and Definition

Some conception and definition need be introduced or given before move on to GDRDC algorithm. First, rough sets theory and its primary features are introduced briefly. Rough set is a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. Pawlak derive the rough probabilities as follow: Defining approximation space $S = \langle U, R \rangle$, where U is a finite nonempty set of research object. $\forall X \subseteq U$, where X called a concept or category on U, $F = \{X_1, X_2, \cdots, X_n\}$ is defined as knowledge on U, $X_i \subseteq U, X_i \neq \phi, X_i \bigcap X_j = \phi$,

$\bigcup_i X_i = U(i \neq j, i, j = 1, 2, \cdots, n)$, R is an equivalence relation on U, $U/R=\{X_1,X_2,\ldots,X_n\}$ denotes a sort of classification, which called a knowledge on U. $[x]_R=\{y \in U \mid xR\ y\}$ denotes the equivalence class of the relation R containing x. every union of elementary sets in S is called a composed set in S, if X is a certain subset of U, then the least composed set in S containing X is called the upper approximation of X, denoted by $\overline{R}(X)$,and the greatest composed set in S contained in X is called the lower approximation of X, denoted by $\underline{R}(X)$. In symbols,

$$\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \phi\}$$

$$\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\}$$

Then boundary region is defined as. $BN_R(X) = \overline{R}(X) - \underline{R}(X)$ On the basis of these forenamed concepts, there are several definitions are given.

Definition 1 (Grid Position-Vector)

Sample $x_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\} \in \mathbf{R}^n$, $x_{ij} \in [t_{j,\min}, t_{j,\max}]$ $(j = 1, 2, \cdots, n)$, suppose dividing interval $[t_{j,\min}, t_{j,\max}]$ to k subinterval and number all from 1 to k, then $g_i = (g_{i1}, g_{i2}, \cdots, g_{in})$ called as position-vector of grid i, where $g_{ij} \in \{1, 2, \cdots, k\}$.

Definition 2 (Grid Density)

For grid $G_i=(G_{i1},G_{i2},\ldots,G_{in})$, if number of sample point contained in it is d, then grid-density of g_i is d.

Definition 3 (Class Kernel)

Suppose the result of initial clustering by some clustering algorithm is $X = \{X_1, X_2, \cdots, X_i\}$, sample space is denote as grid sets $G = \{G_1, G_2, \cdots, G_j\}$, all sample data belong to the same grid G_j are called sample with indiscernibility relation, then the lower-approximation of X_i is called class kernel of X_i, and denoted as $\overline{X} = \{\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_i\}$.

Definition 4 (Freedom Point)

Set sample sets is $U = \{p_1, p_2, \cdots, p_n\}$, n is total sample number. If set $T \subset U, T \cap \overline{X} = \phi, T \bigcup \overline{X} = U$, then all points in T is called freedom point.

3.2 Algorithm Description

Now, we introduce the GDRDC algorithm. The algorithm is mostly based on grid density, rough sets, initial clustering and precise adjustment idea. The algorithm of initial clustering can different, in this paper, we choose SOM as the initial clustering algorithm. By SOM, the given sample data are connected with two-dimensional coordinate neuron by connected weight vector. After the training process of SOM, weight vector is adjusted in neighborhood of each winner so that other units are close to the winner, and the clustering result is get. But it is not easy to cluster factual abnormity shapes or not linearly separable data, in the situation, many clustering method only can get an imprecise clustering result. For improving clustering precision, it is useful to farther adjust by introducing grid density and rough sets idea. The main steps of GDRDC algorithm are given as follow:

Step 1: Set the subinterval number of every dimension k, the total hyperspace grid number is kn, where n is the dimension of sample data. Then determining nonempty grid and sample points number contained in it, which is grid density.

Step 2: Make initial cluster analysis on sample data by SOM, in fact, the initial clustering algorithm can be different, which will be study farther. By SOM, we can get the initial classification. According to definition 5 and 6, the class kernels and freedom point sets can be get, suppose class kernel number is c, freedom point sets contain m point, then they are denoted as $\overline{X} = \{\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_c\}$ and $P = \{p_1, p_2, \cdots, p_m\}$.

Step 3: Make precise adjustment by dynamic clustering method, that is computing mean of Euclidean distance between $p_i (i = 1, 2, \cdots, m)$ and all points contained in $\overline{X}_j (j = 1, 2, \cdots, c)$, the mean is denoted as d_{ij}. For one point p_i, the mean set is denoted as $d_i = \{d_{i1}, d_{i2}, \cdots, d_{ic}\}$. If $\max(d_i) = d_{ij}\ (j = 1, 2, \cdots, c)$, then classify free point i to class j.

Step 4: After step 3, a comparatively precise classification is get. If classification is still dissatisfactory, class abruption or incorporation is need. Step 5 is class abruption method, step 6 is class incorporation method.

Step 5: For every class j, setting standard deviation $\sigma_j = [\sigma_{j1}, \sigma_{j2}, \cdots, \sigma_{jd}]^{\mathrm{T}}$, $\sigma_{ji} = \sqrt{\frac{1}{N_j} \sum_{y_k \in X_j} (y_{ki} - m_{ji})^2}$

Where σ_{ji} is feature i of sample j, y_{ki} is feature i of sample k, $y_k \in X_j$, m_{ji} is feature i of class center j, d is the sample dimension. Setting standard deviation parameter θ_s. Mean distance between the sample in class j and center of class j is denoted as $\overline{\delta}_j$, and distance mean between all sample and every center of class is denote as $\overline{\delta}$. Then compute maximum of standard deviation $\sigma_{j,\max}$ for every class. Suppose for every $\sigma_{j,\max}$, $\exists\ \sigma_{j,\max} > \theta_s$ and meet condition $\overline{\delta}_j > \overline{\delta}$ and $N_j > 2(\theta_N + 1)$, or $c \leqslant K/2$, then class X_j can divide to two classes. Setting initial class center m_j, new class center is m_{j1} and m_{j2}, then new class center can be computed as follow:

Giving k, $k \in (0,1]$, let $\gamma_j = k\sigma_j$, then $m_{j1} = m_j + \gamma_j, m_{j2} = m_j - \gamma_j$. k need make distance

between all samples in X_i to m_{j1} and to m_{j2} different, besides all samples are still in X_i.

Step 6: For all class centers, computing its distance each other $\delta_{ij} = \left\| m_i - m_j \right\| \ (i = 1, 2, \cdots, c-1; i < j)$.

Array them by sort ascending, set incorporation parameter θ_c. Start with the minimum of δ_{ij}, if it less than θ_c, then incorporate them. But when one class *i* need to be incorporated with two class *j* and *k* because δ_{ij} and δ_{ik} is neighboring, it is needed to compute their correlation; two classes with strong correlation are incorporated.

4 INSTANCE ANALYSIS

In order to investigate the efficiency of algorithm we do experiments to compare the algorithm with SOM and k-means algorithm. We choose Iris datasets of UCI data source for clustering experiment. This is the best known database to be found in the pattern recognition literature, the data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two, the latter are not linearly separable from each other[9].

Firstly, we take SOM and K-means algorithm to cluster. When training SOM, we choose learning rate $\eta(t) = 1 - t / epochs$, where *t* is training time, *epochs* is the largest number of training. For the sample data sets is only contains 3 classes, let grid of SOM be 2 ×2 matrix, epochs be 10, 100, 500. Running our programs made by MATLAB 2007. For K-means algorithm, let k=3, using squared Euclidean distances, running by K-means function of MATLAB statistical toolbox. Tab.1 is the results of SOM and K-means algorithm.

Tab. 1 Clustering Results of SOM and K-means

Iris data sets	1-50		51-100			101-150		
	Class 1		Class 2			Class 3		
	Cluster 1	Error (%)	Cluster 2		Error (%)	Cluster 3		Error (%)
SOM (epochs=10)	1-50	0	32 (2)	18 (3)	36	49 (3)	1 (2)	27
SOM (epochs=100)	1-50	0	30 (2)	20 (3)	40	49 (3)	1 (2)	29
SOM (epochs=500)	1-50	0	29 (2)	21 (3)	42	49 (3)	1 (2)	30
K-means	1-50	0	48(2)	2 (3)	23	36(3)	14(2)	38

It can be clearly seen from Tab.1 that SOM and K-means can cluster exactly linearly separable class, but for not linearly separable class, their clustering error is all very big, they classify many sample of one class to another, besides, increasing time of SOM training can't improve the effect.

Now, we use GDRDC algorithm to do experiments on the same data sets. First, let number of subinterval is 4, and then the total number of grid is 256. We can get 30 nonempty grids by computing the position of all samples. Tab. 2 is the density distributing of nonempty grid.

Tab. 2 Density of Distributing of Nonempty Grid

grid	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
density	23	12	6	8	1	3	13	3	16	4	2	1	5	3	3
grid	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
density	2	2	6	5	5	3	5	1	3	1	4	3	5	1	1

Taking SOM to cluster initially, let competition layer structure is 2×3, iterative time is 100. After training, sample data are classified to 5 classes, according to the data of Tab. 2, we can get the class kernel of the five classes, and Tab. 3 is the result.

Tab.3 Class Kernel of Initial Clustering

Class kernel	serial number of nonempty grid	serial number of grid contain freedom point
1	1 2 3 4 5	7 8 9 13
2	6 12 14 16 23	
3	18 20 21 22 24 25 26 27 28	
4	10 11 15 17	
5	19 29 30	

The number of freedom point set is 37. It need to compute respectively the Euclidean Distance mean between every point of freedom point sets with all point of every class kernel, then the point of freedom point set can classify to the class that the Euclidean Distance mean between them is the least. After the step, a more accurate cluster result can be get. Because the classification number of sample data set is smaller than the classification number of initial cluster, we incorporate the most similar classification by step 6 and get the final clustering result, Tab. 4 is the result.

It can be seen from Tab. 4 that the cluster error is lower by using GDRDC clustering algorithm. For class2 and class3, only 3 sample points are classified in error. The

algorithm can determine preferably the need adjust data by SOM, grid density and rough sets idea, then make precise adjustment by dynamic clustering algorithm, finally get a satisfying result. The algorithm excels SOM and K-means clustering algorithm in arbitrary shapes and not linearly separable data sets clustering.

Tab. 4 Clustering Result of GDRDC Clustering Algorithm

Iris data sets		GDRDC cluster result	Error(%)
Class1	1-50	1-50	0
Class2	51-100	51-83 86-100 107	4
Class3	101-150	84 85 101-106 108-150	4

5 CONCLUSION

Along with database technology work up and its application in many fields, a mass of practical data will produce, so clustering analysis technology, an important analysis tool and method of data processing in data mining has become an important research aspect. Now, many clustering algorithms and their improved algorithms are proposed. In this situation, it is a good idea to regard clustering process as two stages consist of rough clustering and precise clustering, following the idea, it is a preferable research aspect to study respectively more applicable algorithm for every process. There is a little attempt in this paper, and a result of instance analysis is satisfying.

REFERENCES

[1] WANG J, ZHOU Z H. Machine learning and its application[M]. Beijing: Tsinghua university publisher, 2006.

[2] JAIN A, MURTY M, FLYNN P. Data clustering: A review[J]. ACM Computing Surveys(CSUR), 1999, 31(3): 264-323.

[3] XU R, WUNSCH D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[4] HAMMOUCHE K, DIAF M, POSTAIRE J G. A clustering method based on multidimensional texture analysis[J]. Pattern Recognition, 2006,39:1265-1277.

[5] VESANTO J. SOM-based data visualization methods[J]. Intelligent Data Analysis 1999,3:111-126.

[6] MA S, WANG T J, TANG S W, etc. A fast clustering agorithm based on reference and density[J].Journal of Software, 2003,14: 1089-1095

[7] ZHANG M L, CHENG Z Q, ZHOU Z H. Survey on SOM algorithm、LVQ algorithm and their variants [J].Computer Science, 2002(29),7:97-100.

[8] AGRAWAL R, GEHRKE J, GUNUPOSUS D. Automatic Subspace Clustering of High Dimensional Data. Data Mining and Knowledge Discovery, 2005,11: 5-33.

[9] BLAKE C L, MERZ C J. UCI Machine Learning repository of machine learning databases. http://mlearn. ics. uci. edu/ MLSummary. Html.