International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Deep Learning and Analyses of Clustering Algorithms

Yang Li

School of Electronic Engineering, Xidian University, Xi'an, 710126, China

Abstract: ***The research actuality and new progress in clustering algorithm in recent years are summarized in this paper. First, the analysis and induction of some representative clustering algorithms have been made from several aspects, such as the ideas of algorithm, key technology, advantage and disadvantage. On the other hand, several typical clustering algorithms an d known data sets are selected simulation experiments are implemented from both sides of accuracy and running efficiency, and clustering condition of one algorithm with different data sets is an analyzed by comparing with the same clustering of the data set under different algorithms. Finally, the research hotspot, difficulty shortage of the data clustering and some pending problems are addressed by the integration of the aforementioned two aspects information. The above work can give a valuable reference for data clustering and data mining.***

Keywords: Clustering algorithm, k-means, Data sets, Refining initial points, Cluster

1. Introduction

Clustering algorithm has been researched for decades, at the same time, clustering is indispensable to research of data mining and pattern recognition. For pattern recognition, clustering mainly used as speech recognition and character recognition. In machine learning, clustering algorithm is applied to image segmentation and machine vision. For image processing, clustering algorithm is applied to data compression and message retrieval, and another important usage of clustering is to apply into data mining, sequence, heterogeneous data analyses, etc.

In this paper, we analyzed clustering algorithm representatively that proposed in recent years about algorithm ideas, key technologies, advantages and disadvantages. And we choose some well-known data sets in experiments, after that, we made conclusion in terms of the analyses[1-3].

Section 1 is introduction of clustering, cluster process and algorithms; section 2 focuses on seventeen representative algorithms; section 3 described eight results of simulation about clustering algorithm, and combined reference [4] for analyses; section 4 is conclusion.

2. Clustering and categories of Algorithm

2.1 Concept of Clustering and Cluster Process

There is no accepted definition of clustering in academia so far. In this paper, we show one definition mentioned by Everitt[5] in 1974: Entities in the same class clusters are analogous, entities in different class clusters are dissimilar; A class cluster is the convergence of the midpoint of the test space, Distance between any two points of the same class cluster is less than that in different class cluster; Class cluster can be described as a regional connectivity which contain high density point set in multidimensional space, they are separated by low-density point set area or other area (class cluster)

As the matter of fact, clustering is an unsupervised classification, and it has no prior knowledge can be used. The following is description of clustering:

Assume we have a set $U=\{p_1, p_2, \ldots, p_n\}$ represents a model set, where p_i is the *i-th* model, $i=\{1,2,\ldots,n\}$; $Ct \subseteq U$, t=1,2,…,k, $C_t=\{p_{t1}, p_{t2}, \ldots, p_{tw}\}$; *proximity*($p_{mx}$, p_{ir}); among them, the first subscript indicates the class which the pattern belongs to; the second subscript indicates a certain mode; function *proximity* is used to describe similarity distance of patterns. If C_t is the result of clustering, it satisfies these following conditions:

1) $\bigcup_{t=1}^{k} C_t = \bigcup$

2) For $\forall C_m, C_r \subseteq U, c_m \neq C_r$, and $C_m \cap C_r = \phi$ (Limited to rigid clustering)

$$J(W,P)=\sum_{i=1,\ldots,k}[\sum_{j=1,\ldots,n} w_{ij}^2 \sum_{m=1,\ldots,t} \lambda_m^r \mid x_{jm}^r - p_{jm}^r \mid^2 + \sum_{j=1,\ldots,n} w_{ij}^2 \sum_{q=t+1,\ldots,m} \lambda_q^c \delta(x_{jq}^c, x_{jq}^c)]$$

Typical clustering processes include data preparation, feature selection and extraction, proximity calculating, clustering (Grouping), making effectively evaluation of clustering results[3,6,7].

Clustering process:

1) Data Preparation: Standardize features and reduce dimensionality.
2) Feature Selection: Choose the most effective feature

from the initial feature, and store it into vector.
3) Feature Extraction: Transform the feature which was selected, in order to shape a new prominent feature.
4) Clustering(Grouping): Choosing a suitable characteristic distance function, measure its closeness, then clustering or grouping.
5) Evaluation of clustering results: Evaluation of external validity; Evaluation of inner validity; Evaluation of the text of correlation.

2.2 Category of Clustering Algorithm

There is no clustering algorithms could be generally used to reveal various structures appeared from multidimensional data aggregation. Clustering algorithm includes several classifications, in this paper, we classify this algorithm into hierarchical clustering algorithm, disconnected clustering algorithm, clustering algorithm based on density and grid, and other clustering algorithm.

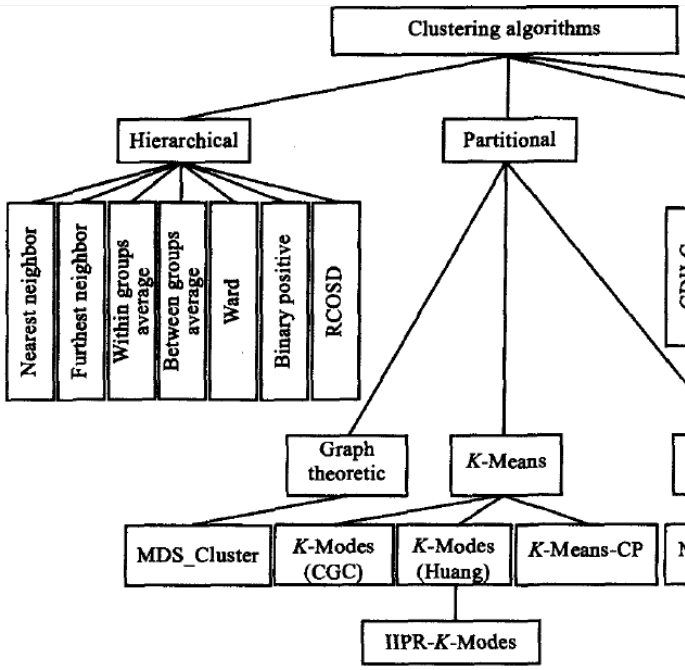

Figure 1: The classification chart of clustering algorithms

3. Clustering Algorithms

3.1 Hierarchical clustering algorithm

Hierarchical clustering algorithm is called Trees clustering algorithm[8,9]. It repeatedly retrievals or clusters data by means of hierarchical structure, in order to form a clustering answer to hierarchical sequence. The complexity of computation is $O(n^2)$, this algorithm is used in classification of small data sets.

Assume a sample set $S = \{O_1, O_2, ..., O_n\}$ has n samples in all

HA1[Initialization]. Regard every sample o_r as a class;

/*form n classes in all: $o_1, o_2, ..., o_n$*/

HA2[Figure out two nearest classes].

$$dis\tan ce(o_r, o_k) = \min_{\forall o_u, o_v \in S, o_u \neq o_v} dis\tan ce(o_u, o_v)\ ;$$

/* Locate the nearest two classes from all classes o_r and o_k*/

HA3[Cluster o_r and o_k]. Then we get a new class o_{rk}; /*Existing class will be reduced by 1*/

HA4. If all the samples belong to a same class, end this algorithm; otherwise, back to **HA2.**

3.2 Traditional clustering rules

Methods of measure the distance between two classes is one important parts of traditional hierarchical aggregation algorithm. In this paper, we use Euclidean distance to measure similarity. Connection rules contain single connection rules, fully connection rules, average connection between class rules, average connection within one class rules, and Ward rules[8](where $||x-y||$ is Euclidean norm, n_i and n_k are the numbers of samples in class o_r and o_k, total methods of two different elements chosen from n_i+n_k denoted as $C(n_i+n_k,2)$);

Single-connection clustering rules:

$$d(o_i, o_k) = \min_{x \in o_j, y \in o_k} \| x - y \|;$$

Full-connection clustering rules:

$$d(o_i, o_k) = \max_{x \in o_i, y \in o_k} \| x - y \|;$$

Average-connection between multiple classes rules:

$$d(o_i, o_k) = (1/n_i n_k) \sum_{x \in o_i} (\sum_{y \in o_k} \| x - y \|);$$

Average-connection within one class rules:

$$d(o_i, o_k) = (1/C(n_i + n_k, 2)) \sum_{x,y \in (o_i, o_k)} \| x - y \|;$$

Ward rules: $d(o_i, o_k) = (1/(n_i + n_k)) \sum_{x \in (o_i, o_k)} \| x - n \|^2$,

where n is the center of fusion clustering.

3.3 New Hierarchical clustering algorithm

(1) Binary-Positive
In 2007, Glebard[4] et al. suggested a new hierarchical clustering algorithm, which is called Binary-Positive. There are many methods represented in Dice to measure various Binary-Positive similarity[10,11].

Glebard et al. adopted these four data sets: Wine, Iris, Ecolic and Psychology balance, to experiment between eleven kinds of clustering algorithms. The results show that all the algorithms are well-used in experiments.

(2) Rough clustering of sequential data(RCOSD)
In 2007, Kumar[12] et al. argued a new hierarchical clustering algorithm based on indistinguishable crude aggregation: RCOSD. This algorithm introduce S^3M as a way to measure similarity. This algorithm could merger more than two classes every time, so it could accelerate the

speed of hierarchical aggregation.

The results show that RCOSD is available, this algorithm could help Webers discriminate potentially significant user groups.

3.4 Partition clustering algorithm

Partition clustering algorithm should specify the number of clusters or center of clusters in advance. By using duplicate iteration calculations, gradually reduce the error of the objective function. When the value of the objective function converges, we get final result of clustering.

3.4.1 *K*-means clustering algorithm

In 1967, Mac Queen mentioned *K*-means clustering algorithm for the first time, so far, several clustering missions have chosen this classical algorithm. The main idea of this algorithm is to find out *K* clustering centers: $c_1, c_2,\ldots,c_k$, in order to minimize every data points x_i and the sum of square of the distance of the closest clustering center c_v(This sum of square of the distance is called deviation *D*).

K-means clustering algorithm[8](clustering of *n* samples)

K1[Initialization]. Assign *K* clustering centers randomly($c_1,c_2,\ldots,c_K$);

K2[Assign x_i]. For every sample x_i, find the closest clustering center c_v, and assign x_i into class c_v that indicated;

K3[Correct c_w]. Put every c_w into indicated class center;

K4[Calculate deviation].

$$D=\sum_{i=1}^{n}[\min_{r=1,\ldots K} d(x_i,c_r)^2];$$

K5[*D* is convergent or not?]. If *D* is convergent, then *return*($c_1,c_2,\ldots,c_K$) and end this algorithm; Otherwise, back to **K2**.

Advantage: *K*-means clustering algorithm could classify large-scale dataset efficiently, and this algorithm calculates faster than Hierarchical clustering algorithm.

3.4.2 *K*-modes algorithm

(1) *K*-modes-Huang algorithm[14]

Introduction of Means and Modes:

In *K*-means algorithm, mean is center of clusters, and can be defined randomly at first. In *K*-modes algorithm, mode is defined as: data set $X=\{X_1,X_2,\ldots X_n,\}$, $\forall X_i\in X$ described by *m* categorical attributes $\{A_1,A_2,\ldots A_m\}$, X_i is denoted as vector $<x_{i1},x_{i2},\ldots,x_{im}>$; *Q* is one of the modes *X*, *Q* is denoted as vector $<q_1,q_2,\ldots,q_m>$, and *Q* satisfies $\sum_{i=1,\ldots,n} d_1(X_i,Q)$ minimum. $d_1(X_i,Q)$ is distance between X_i *and* Q.

Same as *K*-means algorithm, *K*-modes algorithm is also able to produce locally optimal solution, relying on the choice of modes initialization and sequence of centralized data.

In 1999, Huang[15] et al. made the conclusion *K*-means algorithm can only converge local minimum.

(2) *K*-modes-CGC algorithm[16]

In 2001, Chaturvedi et al. suggested a new nonparametric clustering method, which was called *K*-modes-CGC algorithm. This algorithm is similar to traditional *K*-means algorithm(the prior algorithm). *K*-modes-CGC algorithm optimizes one loss function based on norm L_0.

In Monte Carlo simulation, Chaturvedi et al. used *K*-modes-CGC[17] and latent class algorithm to restore a known latent class structure. The results show that, both algorithm have equivalent efficiency.

In 2003, Huang[18] et al. proved that *K*-modes-CGC algorithm is equivalent to *K*-modes-Huang algorithm.

3.4.3 Refining initial points for *K*-modes

In 2002, Sun[19] et al. applied refining initial points for *K*-means suggested by Bradley[20] into *K*-modes(Huang,1998). Sun et al. proposed an experiment based on refining initial points for *K*-modes.

Sun's experiment based on well-known soybean disease dataset. Data of soybean disease contain 47 records, every record was described by 35 characteristics, and every record was signed one of the following four diseases: Diaporthe StemCanker, Charcoal Rot, Rihizoctonia Root Rot, and Phytophthora Rot. Apart from Phytophthora Rot has seventeen records, other three diseases have ten records. There are two programs about *K*-modes:

Program 1: Randomly select initial point sets;

Program 2: Use refining initial points for *K*-modes to select initial point sets.

In addition, program 2 produces precision reliable clustering resultss.

3.4.4 *K*-means Consistency Preservation algorithm (*K*-means-CP)

In 2004, Ding[22] et al. put forward *K*-means Consistency Preservation algorithm. Nearest neighbor consistency is a important concept in statistical pattern recognition, they expended this concept into data clustering: As any data points in one class, its nearest neighbor and mutual nearest neighbor of *k* should belong to this class. They proposed improved algorithm for *k*NN and *k*MN, and regarded the nearest neighbor and mutual nearest neighbor of *k* as an important way of metrics.

K-means-CP:

1[Initialization]. Choose *K* points as centers of the initial class randomly($c_1,c_2,\ldots,c_K$)

2[Assign neighbor sets]. Assign a neighbor set *S*; //*Assign *S* into nearest class C_p,

$$p=\arg\min_{v=1,\ldots,K}\sum_{x_i\in S}(x_i-m_v)^2$$

3[Update clustering center]. Define $m_v=\sum_{x_i\in C_v} x_i/n_v$;

//* Update cluster center (centroid), m_v is center of class C_v, $n_k=|C_k|$

4[Convergent or not?]. If centroid no longer move, then end the algorithm; Otherwise, back to Step2. //* $J_{Km}=\sum_{v=1,\ldots,K}\sum_{x_i\in C_v}(x_i-m_v)^2$ and judge it is convergant or not.

3.4.5 Fuzzy clustering algorithm

In 1969, Ruspini applied fuzzy sets theory into clustering analyses for the first time, and proposed fuzzy clustering means(FCM). FCM is one of the most popular algorithms dealing with image segmentation. FCM can retain more information of initial images. In this paper, we simply introduce the newest study[23,24].

In 2006, Li Jie[25] et al. proposed new algorithm NFWFCA based on feature weighting. Traditional fuzzy *K*-means, *K*-modes, and *K*- prototype all assume that contributions of dimensional feature for each sample vector are the same. In fact, contributions of dimensional feature for each sample vector are discriminate. Based on *K*- prototype, NFWFCA used ReliefF[26] algorithm in order to determine weight of each dimensional feature:

$$\lambda^r=\lambda^r-\frac{diff\_hit^r}{R}+\frac{diff\_miss^r}{R}$$

Property feature weights compute as:

$$\lambda^c=\lambda^c-\frac{diff\_hit^c}{R}+\frac{diff\_miss^c}{R}$$

Revise the object function:

$$J(W,P)=\sum_{i=1,\ldots,k}[\sum_{j=1,\ldots,n}w_{ij}^2\sum_{m=1,\ldots,t}\lambda_m^r\,|\,x_{jm}^r-p_{jm}^r\,|^2+\sum_{j=1,\ldots,n}w_{ij}^2\sum_{q=t+1,\ldots,m}\lambda_q^c\delta(x_{jq}^c,x_{jq}^c)]$$

When $J(W,P)$ is the minimum, the result of clustering is optimal. When $\lambda^c=0$, it corresponds *K*-means; when $\lambda^r=0$, it corresponds *K*-modes; when $\lambda^c\neq 0$, it corresponds fuzzy *K*- prototype.

The results show that this new algorithm is more efficient and accurate. This makes great progress in clustering algorithm researches.

In 2007, Cai[27] et al. combined with local spatial and gray information, proposed clustering algorithm FGFCM based on FCM, characterized: (1) Use a new factor S_{ij} as partial(space and gray) similarity measurement. Not only keep immunity of image, retain image details, but also expect adjustable parameter α; (2) Split time only relate to gray level q, its complexity reduce from $O(NcI_1)$ to $O(qcI_2)$, where c is clustering number, I_1 and $I_2(<I_1)$ are iterations of FCM and FGFCM; (3) FGFCM can be used for many other algorithms, FCM, EnFCM, FGFCM_S1 and FGFCM_S2 all can be deduced.

3.4.6 Graph Theory Algorithm

In 1999, Jain[3] suggested famous graph theory fission clustering algorithm: Construct a minimal spanning tree(MST) based on datas, by deleting the longest leg of minimal spanning tree. The algorithms based on graph theory include Random Walk, CHANMELEON, AUTOCLUST[28,29,30,31].

In 2007, Li[31] suggested a clustering algorithm based on maximum θ distance subtree θ denoted as MDS_CLUSTER. Cut of all edges which length greater than the threshold $\theta\geq 0$, generating maximum θ distance subtree sets, and vertexes of every maximum θ distance subtree make up one class.

4. Clustering Algorithm based on gird and density

Clustering algorithm based on gird and density is one important clustering algorithm, and it is widely used in many sphere especially in spatial information processing.

Different from traditional clustering algorithm, this algorithm could discover arbitrary shape clustering by using data density; this clustering algorithm familiarly combine with other algorithms, especially clustering algorithm based on density.

In 2001, Zhao and Song[32] mentioned a clustering algorithm GDILC which based on grid density contours. The main idea of GDILC is: use density contours image to depict data sample distribution, and use network to calculate each density of data sample. The results show that GDILC has high accuracy and efficient characteristic.

In 2004, Ma[33] proposed a new algorithm SGC which based on shifting grid concept. SGC is a non-parametric algorithm, it does not need users to input parameters, for it divide every latitude of grid structures into one data space. SGC produces displacement concept of whole grid structures, hence, could improve results accurate and efficient.

In 2005, Pileva[34] et al. advanced gird clustering algorithm(GCHL) base on large, high-dimensional database. GCHL combines density-gird clustering algorithm with parallel shaft partitioning strategy, in order to make sure the high density region. This algorithm can be used excellently in random spatial database.

In 2006, Micro[35] et al. faced to moving object trajectory data processing sphere, in view of simple concept of distance between the track, put forward an adaptive clustering algorithm base on density(TFCTMO).

In 2007, Derya[36] et al. expanded DBSCAN(density—based spatial clustering of applications with noise), and then proposed a new algorithm based on density which was called ST-DBSCAN(spatial-temporal DBSCAN), comparing with existing clustering algorithm based on density, ST-DBSCAN has ability in discovering class clusters relying on non-space value, space value, and tense value.

4.1 Other clustering algorithms

4.1.1 ACODF clustering algorithm

In 2004, Tsai[37] proposed a novel with different preferences ant colony system(novel AS)—ACODF(a novel data clustering approach for data mining in large databases), in order to solve clustering problems. ACODF is able to obtain optimal solution quickly, it contain three important strategies as following:

1) The application of different preferences(favorable) ACO strategy. Each ant only visit one tenth of all the cities, then successively reduce number of cities each visit; After several cycles, the concentration of pheromone increased between two closer points, and the concentration of pheromone reduced between two distant points. Therefore, ants would like to visit closer nodes, and use pheromones to strengthen this path, finally, form a high concentration path, then clustering completed.
2) We design two formulas:

$$ns(t+1)=ns(t)\times T$$

where ns is the number of nodes that ants visit in T_0; $ns(t+1)$ is the number of nodes that ants have been visited yet; $ns(t)$ is the number of nodes that ants visited in the last cycle; T is a constant(T=0.95).

$$nf(t+1)=2\times ns(t)/3\text{-}i\times ns(t)/(run\times 3)$$

where nf is the number of nodes that ants visit in T_1; $nf(t+1)$ is the number of nodes that ants have been visited yet; $nf(t)$ is the number of nodes that ants visited in the last cycle; run=2, $i\in\{1,2\}$.

(1) Using tournament selection strategy. Different form traditional ACO, ACODF uses tournament selection strategy in order to select path. That is select K paths randomly in N paths, then choose the shortest path in these K paths($N>K$).

5. Experiments

In this paper, we choose five data sets: Iris, Wine, Soybean, Zoo and Image. Image data set is used to compared with Iris and Wine datasets.

For numerical model data, respectively use Iris, Wine, Image for experiments.

Iris includes 3 classes, each class has 50 elements, and every class represent one kind of flowers, 150 samples-equidistribution in 3 clusters; among them, one class of linear separable with other two class, and other two class are partially overlapped. Data set Wine has good clustering structure, including 178 samples, 13 numerical model property, divided into 3 classes, where contain different amount of samples. Image derived from UCI machine learning data sets, randomly chosen from 7 outdoor image sets.

For categorical attribute data, respectively used Soybean and Zoo dataset for experiments.

Dataset Soybean has 47 samples, including 35 attributes, divided into 4 classes for linear separable. Its attributes are all categorical attributes. Dataset Zoo has 101 records, divided into 7 classes for linear inseparable. In data set Zoo, 16 attributes used to describe samples, 15 of them are Boolean property value {0,1} and 1 categorical attribute(leg count) {0,2,4,5,6,8}.

5.1 Soybean disease data set experiment

We use the formula of accuracy:

$$r=\sum_{i=1,\ldots,k}(a_i/n)$$

Where a_i the sample number which emerges in the i-th cluster(obtain from this algorithm) and initial-points, k is cluster number(k=4), n is total samples number(n=47). Tab.1 and Tab.2 show the experiment result of this algorithm.

Tab.1 Clustering results of 20 random tests for soybean disease data set on 2 algorithms

Cases / Algorithm / Accuracy (%)	*K*-modes	Iterative initial-points refinement *K*-modes
98	5	7
94	6	8
89	0	3
77	0	1
70	7	1
68	2	0

Table 2: Average run time of 20 random tests for soybean disease data set on 2 algorithms

Algorithm	Average running time(s)
K-modes	0.00817331
Iterative initial-points refinement *K*-modes	0.01178265

Basing on the running time, we can figure out that iterative initial-points refinement *K*-modes runs longer than *K*-modes algorithm.

5.2 Hierarchical clustering and *K*-means algorithm

For numerical model data, we randomly did 20 cluster experiments using data set Iris, Wine, Image, and the results show in Tab.4.

Table 4: Clustering results of 20 random tests for Iris, Wine, Image data sets on several algorithms

Algorithm	Average accuracy of running 20 cycles(%)			Average running time(s)		
	Iris	Wine	Image	Iris	Wine	Image
Nearest neighbor	68.00	42.70	30.00	1.5831	3.1346	5.2414
Furthest neighbor	84.00	67.40	39.00	1.5042	3.1434	5.6708
Between groups average	74.70	61.20	37.00	1.5027	3.1526	5.7853
Ward method	89.30	55.60	60.00	2.3793	4.7757	8.9599
K-means	81.60	87.96	56.00	0.0026	0.0038	0.0457

The results show us, operating efficiency of these five clustering algorithms were significantly different between data sets. Thus, in a real world application, we should use different algorithms for different data sets question

5.3 Comparison between *K*-means and *K*-means-CP algorithm

In order to figure out whether *K*-means-CP is obviously better than *K*-means or not, and the relationship between *k*NN consistency and cluster quality, we did 20 random experiments on *K*-means, 1 *K*-means-CP(*k*=1,denoted as cp1), and 2 *K*-means-CP(*k*=2,denoted as cp2), and evaluated the results based on accuracy and quality. The difference within one class, the difference between multiple classes in a whole cluster, and quality can be compute as following formula(1)~(3):

$$\sum_{v=1,\ldots,k}\sum_{x\in C_v} d(x,\overline{x_v})^2 \quad (1)$$

$$\sum_{1\le j<i\le k} d(\overline{x_j},\overline{x_i})^2 \quad (2)$$

$$\sum_{1\le j<i\le k} d(\overline{x_j},\overline{x_i})^2 \Big/ \sum_{v=1,\ldots,k}\sum_{x\in C_v} d(x,\overline{x_v})^2 \quad (3)$$

where k is the number of cluster in clustering result, C_v denotes cluster v, $\overline{x_v}$ denotes the centroid of C_v, $\overline{x_j},\overline{x_i}$ respectively denote the centroid of cluster j and i, d is distance function. Tab.6 shows the result that *K*-means-CP not better than *K*-means, *k*NN consistency has nothing to do with clustering quality.

Table 6: Clustering results of 20 random tests for 5 data sets on *K*-means, cp1&cp2 algorithm

Imagine	Average accuracy (20 times)	Average quality (20 times)
cp1(1NN)	0.623 571 428 571 428	0.778 036 750 839 380 0
cp2(2NN)	0.609 523 809 523 809	0.764 753 617 717 611 0
K-means	0.632 380 952 380 952	0.734 076 358 291 719 0
Iris	Average accuracy (20 times)	Average quality (20 times)
cp1(1NN)	0.840 000 000 000 000	0.258 626 172 448 124 0
cp2(2NN)	0.892 333 333 333 333	0.322 489 157 412 046 0
K-means	0.862 333 333 333 334	0.290 268 692 364 311 0
Wine	Average accuracy (20 times)	Average quality (20 times)
cp1(1NN)	0.898 314 606 741 573	0.045 433 239 324 063 6
cp2(2NN)	0.905 337 078 651 385	0.045 155 360 976 705 9
K-means	0.946 910 112 359 550	0.049 098 735 880 057 5
Glass	Average accuracy (20 times)	Average quality (20 times)
cp1(1NN)	0.511 915 887 850 467	0.400 881 509 658 679
cp2(2NN)	0.531 542 056 074 766	0.404 061 886 906 006
K-means	0.542 523 364 485 981	0.453 522 047 430 905
Ionosphere	Average accuracy (20 times)	Average quality (20 times)
cp1(1NN)	0.691 880 341 880 342	0.003 812 476 851 341 2
cp2(2NN)	0.682 051 282 051 282	0.003 555 311 462 034 7
K-means	0.710 256 410 256 410	0.003 784 599 450 916 1

6. Conclusion

By means of experiments for several clustering algorithm, we can figure out most clustering algorithms need prescribed parameters. Thus, promoting non-prescribed parameters clustering algorithm, combining clustering algorithm with parameters autogeneration algorithm may have good prospect. And RCOSD can efficiently work in

data mining, helping us to understand the results of clustering.

The algorithms discussed in this paper apply to data sets with different properties, accordingly, researchers should use different algorithms and methods for divergent data problems. As a supplement for above conclusion, we compared 11 different algorithms[4] and 8 algorithms proposed in this paper. Tab.7 shows the comparative results of some typical clustering algorithms.

Table 7: Comparative results of several typical clustering algorithms

Algorithm	Years	Sort	Similarity measure	Parameter	Noise	Cluster shape	Scaled dimension	Others
K-means	1967	Partition	Distance function	1	Sensitive	Hypersphere	Large numeric	—
K-means-Huang	1998	Partition	Category similarity function	1	Sensitive	Sphere	Large category	Describe cluster well
K-means-CP	2004	Partition	Distance function	1	Sensitive	Sphere	Large-Scale	*k*NN consistency is irrelevant with clustering accuracy
MDS_CLUSTER	1998	Partition	Eulidean distance	1	In-Sensitive	Arbitrary non-overlap	—	One simple parameter
Feature weighted fuzzy clustering	2004	Partition	Eulidean distance category similarity measure	1	In-Sensitive	Sphere	Small,mix	Feature weighted
Nearest neighbor	1967	Hierarchy	Distance function	1	In-Sensitive	Filamentary	Small and middlow-dimension	—
Furthest neighbor	1967	Hierarchy	Distance function	1	—	Sphere	Small and middlow-dimension	—
Between groups average	1967	Hierarchy	Distance function	1	—	Manifold	Small and middlow-dimension	—
Sequence data rough clustering	2007	Hierarchy	S^3M	2	—	Sequence data	Large-Scale	Depict cluster feature
SGC	2004	Density	Distance function	None	In-Sensitive	Arbitrary shape	Large and middlow-dimension	Mostly used for spatial
GCHL	2005	Grid	Eulidean distance	2	In-Sensitive	Arbitrary shape	Oversize high-dimension	Information processing
ACODF	2004	Others	Eulidean distance	1	—	Sphere，non-sphere	Small, high-dimension	Get optimal value fast

Combining both algorithms in references and the methods proposed in this paper, we can make the conclusion that: clustering algorithms and the results are unpredictable, in practical researches, we should choose appropriate clustering algorithms to solve different kind of data problems in order to obtain the best clustering results.

References

[1] Jain AK, Flynn PJ. Image segmentation using clustering. In: Ahuja N, Bowyer K, eds. Advances in Image Understanding: A Festchrift for Azriel Rosenfeld. Piscataway: IEEE Press, 1996. 65-83.
[2] Cades I, Smyth P, Mannila H. Probabilistic modeling of transactional data with applications to profiling, visualization and prediction, sigmod. In: Proc. of the 7th ACM SIGKDD. San Francisco: ACM Press, 2001. 37-46.
[3] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999, 31(3): 264-323.
[4] Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: An empirical comparison. Data & Knowledge Engineering, 2007, 63(1): 155-166.
[5] Jain AK, Dubes RC. Algorithms for Clustering Data. Prentice-Hall Advanced Reference Series, 1998. 1-334.
[6] Jain AK, Duin RPW, Mao JC. Statistical pattern recognition: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.
[7] Sambasivam S, Theodosopoulos N. Advanced data clustering methods of mining Web documents. Issues in Informing Science and Information Technology, 2006, (3): 563-579.
[8] Marques JP, Written; Wu YF, Trans. Pattern Recognition Concepts, Methods and Applications. 2nd ed., Beijing: Tsinghua University Press, 2002. 51-74.
[9] Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000, 27(4): 305-320.
[10] Zhang B, Srihari SN. Properties of binary vector dissimilarity measures. In: Proc. of the JCIS CVPRIP 2003. 2003. 26-30.
[11] Kumar P, Krishna PR, Bapi RS, De SK. Rough clustering of sequential data. Data & Knowledge Engineering, 2007, 3(2): 183-199.
[12] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tuscon, 1997. 146-151.
[13] Huang Z. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge, Discovery II, 1998, (2): 283-304.
[14] Huang Z, Ng MA. Fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans. on Fuzzy Systems, 1999, 7(4): 446-452.
[15] Chaturvedi AD, Green PE, Carroll JD. *K*-modes clustering. Journal of Classification, 2001, 18(1): 35-56.
[16] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61(2): 215-231.
[17] Huang ZX, Michael K. A note on *K*-modes clustering. Journal of Classification, 2003, 20(2): 257-261.
[18] Sun Y, Zhu QM, Chen ZX. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, 2002, 23(7): 875-884.
[19] Bradley PS, Fayyad UM. Refining initial points for *k*-means clustering. In: Proc. of the 15th Internet Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1998. 91-99.
[20] Ding C, He X. *K*-Nearest-Neighbor in data clustering: Incorporating local information info global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584-589.
[21] Lyer NS, Kandel A, Schneider M. Feature-Based fuzzy classification for interpretation of mammograms. Fuzzy Sets System, 2000, 114(2): 271-280.
[22] Yang MS, Hu YJ, Lin KCR, Lin CCL. Segmentation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithm. Journal of Magnetic Resonance Imaging, 2002, (20): 173-179.
[23] Li J, Gao XB, Jiao LC. A new feature weighted fuzzy clustering algorithm. ACTA Electronic Sinica, 2006,34(1): 412-420.
[24] Kononenko I. Estimating attributes: Analysis and extensions of relief. In: Proc, of the 17th European Conf. ON Machine Learning. LNCS 784, 1994. 171-181.
[25] Cai WL, Chen SC, Zhang DQ. Fast and robust fuzzy *c*-means clustering algorithms incorporating local information for image segmentation. Pattern Recognition, 2007, 40(3): 825-833.
[26] Harel D, Koren Y. Clustering spatial data using random walks. In: Proc. of the 7th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining. New York: ACM Press, 2001. 281-286.
[27] Karypis G, Han EH, Kumar V. CHANELEON: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer, 1999, 2(8): 68-75.
[28] Estivill-Castro V, Lee I. AUTOCLUST: Automatic clustering via boundary extraction for mining massive point-data sets. In Abrahart J, Carlisle BH, eds. Proc. of the 5th Int'l Conf. on Geocomputation. 2000. 23-25.
[29] Li YJ. A clustering algorithm based on maximal θ-distant subtrees. Pattern Recognition, 2007, 40(5): 1425-1431.
[30] Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of the Internet Conf. on Info-Net. Beijing: IEEE Press, 2001. 140-145.
[31] Ma WM, Chow E, Tommy WS. A new shifting grid clustering algorithm. Pattern Recognition, 2004, 37(3): 503-514.
[32] Pilevar AH, Sukumar M. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern Recognition Letters, 2005, 26(7): 999-1010.
[33] Nanni M, Pedreschi D. Time-Focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems, 2006,27(3): 267-289.
[34] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[35] Tsai CF, Tsai CW, Wu HC, Yang T. ACODF: A novel data clustering approach for data mining in large databases. Journal of Systems and Software, 2004, 73(1):133-145.