ATLANTIS PRESS

Advances in Computer Science Research, volume 87

3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

Research on Density Sensitive Clustering Algorithm for Non-convex Sets

Liwen Song [a], Jiahui Qi [b], Min Wu [c]

Center of Modern Educational Technology, University of Science and Technology of China Hefei, China

[a]slw523@mail.ustc.edu.cn, [b]jhqi@ustc.edu.cn, [c]minwu@ustc.edu.cn

Abstract. Applying Clustering to non-convex data is a challenging task, and traditional clustering algorithms often fail to achieve good results. In this paper, an improved spectral clustering algorithm based on density sensitivity (DSISC algorithm) is proposed. By using the ensemble selection strategy for the mean shift algorithm, relatively good optional clusters are selected from the non-convex data sets, and then the number of clusters is transported into the spectral clustering algorithm as input, and the density-sensitive distance is used as the similarity measure. The experimental results give us clear information that the DSISC is better than traditional mean shift algorithm and spectral clustering algorithms in normalized mutual information clustering error rate.

Keywords: mean shift, spectral clustering, density sensitivity, ensemble selection.

1. Introduction

Many classic clustering algorithms (represented by k-means algorithm [1]) have good performance in convex data sets, but when the datasets are non-convex sets, not spherical distributions, or when the data points overlap seriously, the performance of the algorithm is poor, such as the two circle data sets with double rings, the traditional clustering algorithm can hardly get accurate results. In addition, the traditional algorithm uses the iterative optimization method to find the optimal solution so it cannot guarantee the solution convergence to the global optimal solution [2].

Spectral Clustering (SC) algorithm [3] is a fashion and effective algorithm which takes graph theory as its fundamental. Compared with the traditional clustering algorithm, the spectral clustering algorithm has obvious advantages. It is not only simple to implement, but also can identify any type of sample space, then converge the solution to the global optimal solution. It is very suitable for practical problems, such as image segmentation [4].

However, although the spectral clustering algorithm has achieved good results, there are still many problems worthy of further study, such as the need choose the number of final clusters as input parameters, such as the determination of sample similarity measure in spectral clustering, besides, spectral clustering also has problems such as standardization of Laplacian matrix, selection of Laplacian matrix feature vectors and high computational complexity, parallelization of large data, nonnegativity and sparseness of clustering matrix, and so on.

To gauge the similarity between samples, we introduce density sensitivity. The common clustering algorithm uses Euclidean distance as the similarity measure, which often fails to get satisfactory clustering results for most clustering problems with complex structures. In order to reflect the actual spatial distribution characteristics of data, Wang Ling et al. proposed a density sensitive distance [5], and then applied it to spectral clustering algorithm with good results.

Considering the shortcoming that spectral clustering algorithm needs to manually determine the number of clusters, we introduce mean shift (MS) algorithm [6], which is a method based on nonparametric density estimation. MS algorithm does not attach any assumptions to the data distribution law, but studies the data distribution characteristics directly from the data samples themselves, requires little prior knowledge, relies entirely on training data to estimate, and can handle any probability distribution. By using MS algorithm, we get the number of the cluster as a candidate for algorithm input into subsequent spectral clustering.

In this paper, we propose an improved spectral clustering algorithm based on density sensitivity (DSISC algorithm), which consists of two key modules: (1) module of ensemble selection for MS algorithm (2) module of density-sensitive spectral clustering.

2. Relevant Knowledge

2.1 Clustering Algorithm for the Non-Convex Dataset

If the points on the line between any two points in the collection are all in the collection, the collection is called a convex set. Datasets that do not meet the definition of convex sets are non - convex. On the optimality level, non-convex means that going all the way along the gradient direction can only guarantee the local optimum, not the global optimal solution. At present, a large number of clustering algorithms have a very good performance on convex sets, but in practice, there are many cases where datasets are non-convex sets, such as double-ring data sets or data sets with serious data overlap, because the data in practical problems are not necessarily evenly distributed or even in density, the distance between clusters may be very long. At this time, if we still select some traditional clustering algorithms, the clustering results are often not satisfactory.

However, in the traditional clustering algorithm of machine learning, there are also a few algorithms that can cluster non-convex data sets, such as DBSCAN algorithm[7], which is a very typical density clustering algorithm. Compared with K-Means and BIRCH, DBSCAN can be applied to both convex and non-convex datasets, and can cluster dense data sets of any shape and find noise point while clustering. However, DBSCAN also has many disadvantages. (1) If the density of the sample set is uneven and the distance between clusters is too long, the cluster quality is poor. (2) Parameter adjustment is complex, which mainly requires select distance threshold and neighborhood sample number threshold MinPts at the same time. Different parameter combinations have a great influence on the final clustering effect. (3) The larger the sample set, the longer the clustering convergence time. In addition to DBSCAN algorithm, we know that in the new clustering method, spectral clustering algorithm has a good effect on clustering non-convex set data.

2.2 Spectral Clustering

Compared with some classic clustering algorithm, the spectral clustering algorithm can identify non-convex data sets. Spectral clustering algorithm takes the similarity matrix W between samples as fundamental, then finds out the internal relationship between data points by calculating the feature vector. The algorithm does something with the number of data points, and is not related to the dimension. this is because spectral clustering only needs the similarity matrix between data, so it is very effective for processing sparse data clustering, thus spectral clustering can avoid the singularity problem caused by high-dimensional feature vectors.

3. Improved Spectral Clustering Algorithm based on Density Sensitivity

3.1 Ensemble Selection for MS Clustering Module

(1) Improvement of spectral clustering process – a combination of MS and SC

The reason why the Mean shift(MS) clustering algorithm was chosen is mainly because (1) MS algorithm is a method based on nonparametric density estimation, which does not attach any assumptions to the data distribution law, but studies the data distribution characteristics directly from the data samples themselves, (2) MS algorithm requires little prior knowledge, relies entirely on training data to estimate, and can handle any probability distribution problem.

The reason (1) is that the MS algorithm is very suitable for irregular non-convex data sets and is no longer constrained by the data sets. The reason (2) just avoids one of the defects of SC algorithm, that is, the number of final clusters is required as the input parameter of the algorithm, which will lead to the following problems: (1) the parameters need to be adjusted continuously. (2) it is difficult for us to have an accurate grasp of the number of final clusters for large-scale data sets. The above

two questions will not only cost us a lot of time but also increase the difficulty of our experiment. In view of the above problem that SC algorithm needs to consider, let's not cluster the data directly through SC algorithm, but select MS algorithm to process the data set in the first module of DSISC algorithm. Our expectation is to obtain the number of the cluster as a candidate through MS algorithm and deliver this value to the second module of DSISC algorithm for calculation of SC algorithm.

During the implementation of MS algorithm, we define R^d, which represents a d-dimensional Euclidean space, the sampling points in the space are $\{x_i, 1 \le i \le n\}$, we set the weight of sampling points equal, that is $w(x_i) = 1/n$, the bandwidth matrix is proportional to the unit matrix $H_i = h^2 I$, and the mean shift iterative formula is:

$$x_{i+1} = m_{h(x)} = \frac{\sum_{i=1}^{n} g(\| (x - x_i) h^{-1} \|^2) x_i}{\sum_{i=1}^{n} g(\| (x - x_i) h^{-1} \|^2)} \tag{1}$$

The bandwidth h is the radius of circle O which represents the search area, and the center of the circle is continuously moved to the high-density area by calculating the drift vector from all the data points in the search area circle O to the center of the circle. However, in this process, we found that the determination of h value also has a great influence on the result. In order to reduce the influence of h value on the final result, we introduced the ensemble selection strategy [8].

(2) Improvement on MS clustering results processing—Ensemble selection strategy

Clustering Ensemble refers to combining multiple clustering results into a final division of a given task. It relates to two main steps: (1) the generation of cluster members; (2) the final division of cluster is obtained by calculating cluster members through consensus function [9]. After obtaining the clustered members, we also have an important step, that is, selecting the clustered individuals, we need to take the generated cluster individuals as new data points and resampling them. Each time a certain number of MS cluster individuals are selected for cluster ensemble, an ensemble solution is obtained, and then the correlation between each cluster individual and the ensemble solution is calculated. Let Grade indicate this correlation and define it as follows

$$Grade = 1 - NMI(P_i, P^*) \tag{2}$$

or

$$Grade = 1 - ARI(P_i, P^*) \tag{3}$$

among them represents clustering individuals, is the ensemble solution of partial cluster individuals. The process of sampling will be taken several times and cluster individuals were selected based on the combined Grade values.

Define the number of generated MS cluster individuals as H, and the set of MS cluster individuals is $\prod = \{\pi_1, \pi_2, ... \pi_H\}$, $GC^{(k)} = (Grade^k(\pi_1), Grade^k(\pi_2), ..., Grade^k(\pi_H))$ represents the Grade value vector of the kth individual sampling. We assume that we have T clusters individual Grade value $(GC^{(1)}, GC^{(2)}, ...GC^{(T)})$, then we calculate the final merged Grade value $GC^{(final)}$, all member individuals are sorted by $GC^{(final)}$, and select the number of required individuals according to the sorted value. One of the important steps is to combine the Grade values from multiple sampling. Given multiple Grade values, the combination function Γ is defined as a mapping that maps all T Grade values to a consensus value:

$$\Gamma : \{GC^{(k)} \mid k \in \{1, 2, ..., T\}\} \to GC^{(final)} \tag{4}$$

$GC^{(k)} = (Grade^k(\pi_1), Grade^k(\pi_2), ..., Grade^k(\pi_H))$; $Grade^k(\pi_j)$ is the value of the cluster individual π_j in the Grade solution $GC^{(k)}$, where $k \in \{1,2,...,T\}$. The combination process is to find a consensus solution that satisfies the following formula

$$GC^{(final)} = \arg\max_{GC} \sum_{k=1}^{T} \| GC^{(k)}, GC \| \tag{5}$$

$\| \circ \|$ is a measure of similarity measure between two Grade values. The Euclidean distance here is used to measure the similarity between the two Grade values. Therefore, the objective function is defined as maximizing the following function $F(GC)$, That is:

$$F(GC) = \sum_{i=1}^{T} \| GC - GC^{(i)} \|^2 \tag{6}$$

After calculating the derivation of the above formula, we get the following result:

$$\frac{dF}{dGC} = 2\sum_{i=1}^{T} (GC - GC^{(i)}) \tag{7}$$

Set the above formula equal to 0, and the value of GC can be obtained as follows:

$$GC = \frac{\sum_{i=1}^{T} GC^{(i)}}{T} \tag{8}$$

Therefore, we can use the average Grade value as a measure of the effectiveness of clustering individuals and select individuals with larger Grade values as members of clustering individuals. This is the complete process of ensemble selection for MS clustering.

3.2 Spectral Clustering Module based on Density Sensitivity Module

(1) Density sensitivity

The classic Euclidean distance can only reflect the local consistency feature of the cluster structure, but it is not able to reflect the global consistency feature. To reflect characteristics of the data's the actual spatial distribution, Wang Ling et al. proposed a density-sensitive distance, and then applied it to spectral clustering algorithm with good results.

Define 1 Density Adjustable Line Length:

$$L(x, y) = \rho^{dist(x,y)} - 1 \tag{9}$$

In the formula $dist(x, y)$ represents the Euclidean distance between data points x and y, $\rho > 1$ is known as a scaling factor. The line length defined as above can be used to describe the global consistency of clustering. We enlarge or shorten the length of the line between two points by adjusting the scaling factor ρ. Based on the adjustable length of a line segment, a density-sensitive distance measure can be further defined.

Definition 2 sets of data points $V = \{x\}_{i=1}^{n}$ as the vertex of the graph $G(V, E)$ .Set $\rho = \{p_1 \to p_2 \to ... \to p_l\} \in V^l$ as the path between connecting vertex p_1 and connecting vertex p_l, $l = |\rho|$ is the number of the vertices, where the edges (p_k, p_{k+1}), p_{ij} ,$1 \le i$, $j \le n$ indicates the set of all the paths between x_i and x_j ,so the density sensitive distance between x_i and x_j is defined as

$$D_{ij} = \min_{\rho \in p_j} \sum_{k=1}^{l-1} L(p_k, p_{k+1}) \tag{10}$$

$L(\circ,\circ)$ represents the length of a line segment with adjustable density between two points. Therefore, we can construct the similarity matrix $W \in R^{n\times n}$ according to the density-sensitive similarity measure,

Of which

$$W_{ij} = \frac{1}{D_{ij} = \min_{\rho\in p_j} \sum_{k=1}^{l-1} (\rho^{dist(p_k, p_{k+1})} - 1) + 1}, W_{ii} = 0 \tag{11}$$

(2) Spectral clustering based on density sensitivity

Firstly, the data set is regarded as the node set V of the graph, the similarity between data points represents the weights of edges, and the similarity between all data points constitutes the edge set E; Then the data set is constructed into a graph $G(V,E)$, and the adjacency matrix of the graph is taken as a similarity matrix and recorded as $M \in \mathbb{R}^{m\times n}$:Then add up elements of every column in W to get n number (n is the number of nodes), use them set the values of diagonal (all other places are 0), and a new Degree Matrix is formed, denoted as D . Therefore:

$$d_i = \sum_{j=1}^{n} w_{i,j} \tag{12}$$

set $L = D - W$, here Laplacian Matrix is L. The first k eigenvalues of L and the corresponding eigenvectors need to be calculated. Finally, the k eigenvectors are arranged together to construct a $n \times k$ matrix, regard every row as a vector in the k-dimensional space (thus dropping from the high-dimensional space to the low-dimensional space), and clustering is carried out by using k-means algorithm, and each row in the clustering result belongs to the class to which the node (i.e. the initial n data points) in the original graph g belongs respectively.

3.3 DSISC Algorithm

In this paper, DSISC algorithm is proposed. The algorithm takes the advantages of MS and SC algorithm and takes them as the starting point, non-convex set data is its mission target. The improvement is proposed for the shortcomings of the two algorithms. DSISC is mainly divided into Ensemble selection for MS Clustering Module and Spectral Clustering Module Based on Density Sensitivity Module

4. Experimental Results and Analysis

4.1 Basic Information of UCI Data Set used in the Experiment

Table 1. DATASET Information

Dataset Name	Number of instances	Attribute	Category ratio
Ionosphere	351	34	1:1.8
Hepatitis	155	19.	1:4
Pima Indian Diabete	768	8.	1:2
Sonar	208	60	1:1.2
Heart-Statlog	270	13	1:11

Algorithm: DSISC algorithm

Input: dataset $X = \{x_1, x_2, ..., x_n\}$, bandwidth parameter σ, dimension P after dimension reduction by spectral clustering

Output: Cluster Collection $C = \{C_1, C_2, ..., C_k\}$

First, Ensemble selection for MS Clustering Module

1) In MS algorithm, randomly select bandwidth parameters σ_i in $[\sigma_{\min}, \sigma_{\max}]$ in the cluster obtain cluster individuals $\prod = \{\pi_1, \pi_2, ...\pi_H\}$;

2) Randomly select $\left[H/2 \right]$ individual cluster members to generate a subset of cluster individuals;

3) Use $\Pi^{(k)}$ to get a consensus division $Con^{(k)}$.

4) Calculate $Con^{(k)}$ and the correlation of each cluster individual member to obtain

5) Calculate $GC = \frac{\sum_{i=1}^{T} GC^{(i)}}{T}$, sort $GC^{(final)}$, and select the first Num cluster individuals as the chosen cluster individuals

6) Merge subsets of cluster individuals by consensus function to obtain a final consensus partition, i.e. the number of the cluster as a candidate

Second, Spectral Clustering Module Based on Density Sensitivity Module

1) Construct adjacency matrix W and degree matrix D according to equation (11) (12)

2) Calculate Laplace matrix L and construct a standardized Laplace matrix $D^{-1/2}LD^{-1/2}$

3) Calculate the eigenvectors f corresponding to the smallest p eigenvalues in $D^{-1/2}LD^{-1/2}$.

4) Normalize the matrix composed of the respective corresponding feature vectors f according to the rows to finally form the $n \times P$ dimension feature matrix F

5) Take each row in F as a p-dimensional sample, n samples are clustered by k-means clustering method

4.2 Comparative Experiment and Parameter Setting

The DSISC algorithm is compared with the following two algorithms, and the experiment is repeated 20 times for each dataset, and then the average value is taken as the final experimental result of different algorithms in each data set.

(1) SC algorithm: use standard Gaussian function as the similarity between data points:

$$W(x_i, x_j) = \exp(-\| x_i - x_j \|^2 / 2\sigma^2)$$

Search each scale parameter σ in $\{4^{-1}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0\}$, σ_0 is the average distance between any two data points in the dataset.

(2) KASP algorithm: KASP algorithm (k-means-based approximate spectra clustering) firstly applies k-means algorithm to the data set to get k representative points of the data set, then divide all data points closest to the representative points into one group, and finally apply spectral clustering algorithm to each group. Here, the parameter k in KASP algorithm is set at $\{2k, 4k, ...20k\}$

4.3 Experimental Results and Analysis

Evaluation: Normalized Mutual Information (NMI) and Clustering Error (CE)[10].Specific definitions are as follows:

(1) Normalized Mutual Information (NMI):

$$NMI(x, y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}$$

Where X and Y are two random variables, $I(X,Y)$ is the mutual information of X and Y, H(X) and H(Y) are the entropy of X and Y. After the normalized mutual information is used in the clustering performance evaluation, the formula becomes:

$$NMI = \frac{\sum_{l=1}^{|C|}\sum_{k=1}^{|C|} n_{l,k} lb\left(\frac{nn_{l,k}}{n_l \hat{n}_k}\right)}{\sqrt{\sum_{l=1}^{|C|} n_l lb\frac{\hat{n}_k}{n}}}$$

Where n_l represents the number of data points contained in the cluster $C_l (1 < l <| C |)$ after the cluster is completed, $\hat{n}_k$ indicates the number of data points actually contained in the $h(1 < l <| C |)$ th class; $n_{l,k}$ represents the number of data points that cluster C_l and hth class contained at the same time. The larger NMI, the better the clustering effect.

(2) Cluster Error Rate (CE)

$$CE = 1 - \frac{\sum_{i=1}^{n} \delta(\hat{c}_i, map(c_i))}{n}$$

Where $\hat{c}_i$ and c_i respectively represent the actual class labels of the data points and the class labels obtained after clustering; $\delta(x,y)$ is an incremental function, i.e. $\delta(x,y) = 1$ when x=y, otherwise $\delta(x,y) = 0$; Map() is a mapping function that maps the class labels obtained after clustering to the actual class labels. Obviously, the smaller the CE value, the better the performance of the clustering algorithm.

Table 2. Comparison of NMI and CE Indices of Algorithms

dataset	Evaluation index	SC	KASP	DSISC
Ionosphere	NMI	0.2504	0.3303	0.4258
	CE	0.3177	0.2831	0.1800
Hepatitis	NMI	0.5813	0.6867	0.7750
	CE	0.4495	0.3505	0.2333
Pima Indian Diabetes	NMI	0.4589	0.5828	0.6800
	CE	0.4222	0.4978	0.2169
Sonar	NMI	0.5626	0.6879	0.7696
	CE	0.3824	0.2174	0.1850
Heart-Statlog	NMI	0.4407	0.5444	0.5750
	CE	0.3893	0.2319	0.2211

In the experiment, SC algorithm, KASP algorithm and DSISC algorithm were run on 5 datasets respectively, and their NMI and CE were recorded in Table 1. Through data analysis, it can be clearly seen that SC and KASP algorithms using Euclidean distance as similarity measure are basically higher in CE error rate than DSISC algorithm using density sensitive distance as similarity measure, and the NMI of DSISC is also significantly higher than the NMI of the other two algorithms. It can be seen that Ensemble selection for MS Clustering Module in DSISC algorithm improves clustering effect obviously.

5. Conclusion

The improved spectral clustering algorithm based on density sensitivity (DSISC) proposed in this paper divides the non-convex dataset clustering problem into two layers and gradually clusters them, which is a divide-and-conquer clustering idea. DSISC algorithm makes full use of the advantages of MS and SC algorithm, so it is more effective for clustering large and complex datasets. DSISC algorithm uses the density-sensitive distance as the similarity measure between data, thus more accurately it describes the actual distribution characteristics of sample data. Experimental results show that DSISC algorithm performs better than KASP algorithm and ordinary spectral clustering algorithm when the data set is large and complex.

References

[1]. Macqueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]// Proc of Berkeley Symposium on Mathematical Statistics & Probability. 1965.

[2]. Duda R O, Hart P E, Stork D G. Pattern classification /[M]// Pattern classification. 2001.

[3]. Dhillon I S, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts[C]// Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2004.

[4]. Malik J, Belongie S, Leung T, et al. Contour and Texture Analysis for Image Segmentation[J]. International Journal of Computer Vision, 2001, 43(1):7-27.

[5]. WANG Ling, BO Lie-feng, JIAO Li-cheng. Density-Sensitive Spectral Clustering[J]. ACTA ELECTRONICA SINICA, 2007, 35(8):1577-1581.

[6]. Fukunaga K, Hostetler L D. The estimation of the gradient of a density function, with applications in pattern recognition[J]. IEEE Trans. inf. theory, 1975, 21(1):32-40.

[7]. Viswanath P, Pinkesh R. l-DBSCAN: A Fast Hybrid Density Based Clustering Method[C]// International Conference on Pattern Recognition. 2006.

[8]. Fred A. Finding Consistent Clusters in Data Partitions[M]// Multiple Classifier Systems. 2001.

[9]. Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining partitionings [J]. Journal of Machine Learning Research, 2002, 3(3):583-617.

[10]. Wu M. A local learning approach for clustering[C]// International Conference on Neural Information Processing Systems. 2007.