An Effective Clustering Method Based on Shared Nearest Neighbors and Graph

Ruiqi Han

Shenzhen Middle School, Guangdong, Shenzhen 518000, P.R.China

Abstract
Clustering is of central importance in pattern recognition, machine learning and engineering fields. However, the main drawback of many existing methods is that the optimal parameter values have to be specified in advance in these clustering algorithms, and the clustering results are highly dependent on the assignment of parameters. Generally, it is difficult to set the optimal parameter because of the lack of prior knowledge about classification information in unsupervised learning. To deal with this problem, this work proposes a nonparametric clustering method (NPCM) to partition dataset into subsets based on shared nearest neighbors. The key idea of this approach is to construct a graph according to the sharing nearest neighbors of two samples. In the constructed graph, each sub graph represents a cluster. Unlike most existing clustering algorithms, no parameter is used in the proposed method. For datasets with well-separated classes, the proposed method can effectively detect the clusters, regardless their density, size and shape. A large number of experimental results confirm the effectiveness of the method.

Keywords: pattern recognition, clustering, shared nearest neighbors, graph,

1. Introduction
Clustering is an unsupervised learning method that attempts to find some patterns in unlabeled datasets. It is one of the important topics in pattern recognition, data mining, and machine learning. Generally, clustering is to divide the dataset into several subsets, such that the samples in the same class are similar in a sense, and the members of different classes are dissimilar in the corresponding sense. An investigation of current clustering algorithms existed in the open literatures make it clear that cluster detection is still an experimental science. The performance of a clustering algorithm is usually not only dependent upon the type of data set, but is also strongly influenced by the chosen measure of sample similarity, the assignment of parameters, the distribution and the density of the dataset [1,2,3]. Therefore, it is still a challenging task to develop new and effective clustering approach.
A large number of clustering algorithms have been proposed for artificial or real datasets in past decades. The most famous and classical cluster algorithms should be c-means, fuzzy c-means (FCM) [4] and density-based spatial clustering of applications with noise (DBscan)[5]. Usually, the desired cluster result can be obtained successfully for the datasets with spherical clusters by c-means and FCM algorithms if the number of classes is assigned correctly in advance. In general, it is difficult to know the ideal number of classes previously since clustering is an unsupervised classification process. To deal with this problem, many different cluster validity indices have been proposed based on FCM algorithm in past decades [6,7], aiming at finding a set of clusters that best fit natural partitions without any class information. However, because the sample is always assigned to the nearest center, these approaches are not able to detect non-spherical clusters if Euclidean distance is adopted.

For the clusters with arbitrary shapes, DBscan algorithm exhibits its excellent classification ability if two parameters, Minps (minimal points) and radius, are set properly [5]. However, DBscan has not a good performance when the clusters have large variation in density. Furthermore, DBScan is sensitive to its parameters, and different parameter settings are likely to produce very different cluster results. To overcome some of these shortcomings, several improvements of DBscan were introduced. OPTICS algorithm can well group the samples with different densities and shapes [8]. EnDBscan algorithm is also an efficient variation of DBscan [9].

Clustering based on shared nearest neighbors (SNNs) was first introduced by Jarvis and Patrick [10]. A nonparametric clustering technique was suggested which combined the concept of similarity with nearest neighbor sharing. Although they declared that their method was a nonparametric clustering technique, two parameters were still used in their algorithm. By using of nearest neighbors (NN), Ritter et al [11] presented a method for autonomously finding clusters, eliminating background noise, outliers, and detection of clusters with different densities on a given data set. Based on the near neighbor influence, Chen [12] introduced a clustering algorithm CNNI, an improved version in time cost of CNNI algorithm ICNNI, and a variation of CNNI algorithm VCNNI.

The k-nearest neighbors (kNN) is one of the most fundamental and powerful methods in data mining and pattern recognition. As a basic technique, it has been widely used in a number of clustering or classification methods. The typical application should be spectral clustering method which uses the eigen values of the similarity matrix of the data to obtain its clusters [13]. The key problem in spectral clustering algorithms is to construct the affinity graph which affects directly the cluster result. Based on kNN method, Liu et al [14] studied the problem of how to reliably compute neighborhoods on affinity graphs and introduced a kDN based clustering method. Hastie and Tibshirani [15] proposed a locally adaptive form of nearest neighbor classification to try to ameliorate this curse of dimensionality.
Several novel clustering algorithms have recently been proposed in machine learning [16,17,18]. One is the affinity propagation algorithm, which takes the similarity between data points as the input measure [19]. The synchronization clustering method tried to find the intrinsic structure of the data without any distribution assumptions by using a dynamic synchronization process [20]. Rodriguez and Laio designed a magical clustering algorithm by search and find of density peaks [21].

Most of clustering algorithms published recently focus on improving some famous clustering algorithms or extending some clustering algorithms by various techniques [22-25]. Little attention has been paid to the nonparametric cluster for unsupervised learning. Several nonparametric cluster methods are investigated by the Gaussian process and the Dirichlet process [26,27], maximum margin frame [28] and nonparametric density estimator [29]. However, the parameters or thresholds are still used in ref.[26-29].

It should be noted that it is impossible to know the best parameter values or thresholds for a given data set in advance. Usually, the best parameter values are experimental values. Furthermore, the cluster results are varied with the different parameter values.

Motivated by the kNN, DBscan and SNNs methods, we in this paper propose a simple nonparametric cluster method based on shared nearest neighbors. Unlike all the existed clustering algorithms, no parameter is used in our proposal. The clustering results can be obtained by the construction of the connected sub graphs. The clusters with different shape, density and size can be discovered by the proposed method. Extensive experimental results on several classical artificial data sets and real data sets demonstrate the validity and practicability of our method.

2. Nonparametric clustering method based on shared nearest neighbor

Let \( DS=\{x_1, x_2, \ldots, x_n\} \) be a dataset in an \( m \)-dimensional space. Naturally, it is required that these \( n \) samples be partitioned into \( C \) clusters. Generally, the samples in the same cluster should have more similarity than those in different clusters. From the point of graph, this indicates that there are more dense connections within clusters than cross-class links. If we construct the links among these samples properly, the desired sub graphs will be obtained. Thus, satisfactory cluster results can be found.

Based on shared nearest neighbors, we would like to construct the affinity graph by following steps.

Step 1: (Finding nearest neighbors) For each sample \( x_i \in DS \), compute the Euclidean distance between \( x_i \) and \( x_j \in DS, j=1, 2, \ldots, n, j\neq i \), and arrange them in an ascending order. Denote the nearest neighbors of \( x_i \) by \( N_k(x_i) = (x_{i,1}, x_{i,2}, \ldots, x_{i,k}) \), where \( x_{i,s} \) represents the \( s \)-th nearest neighbor of \( x_i \).
Step 2: (Initial linking) Link $x_i$ to its three nearest neighbors. For the first nearest neighbor $x_{i,1}$ of $x_i$, consider the intersection $\text{SNN}_3 = N_3(x_i) \cap N_3(x_{i,1})$. If $\text{SNN}_3 = \emptyset$, then compute $x_{i+1}$.

Step 3: (Expanding shared nearest neighbors) From $k=4$ to $n$, judge orderly whether or not $x_{i,k}$ is in $N_k(x_{i,1})$. If $x_{i,k}$ is not in $N_k(x_{i,1})$, then goto step 4. Otherwise, it implies that $x_{i,k}$ is one of the shared nearest neighbors of $x_i$ and $x_{i,1}$. To keep the sparse connections of inter-clusters, compute the intersection $\text{SNN}_{k-1} = N_{k-1}(x_i) \cap N_{k-1}(x_{i,k})$. If $\text{SNN}_{k-1} = \emptyset$, go to step 4; else set $k = k+1$.

Step 4: (Linking) Connect $x_i$ to each sample in $N_{k-1}(x_i)$ and then compute $x_{i+1}$.

It is possible that $x_i$ will link all the rest samples in DS, if $x_i$ and $x_{i,1}$ are so near that all the remaining samples in the DS are the nearest neighbors they share. In this case, we can safely conclude that $x_i$ and $x_{i,1}$ are in the same cluster and simply link $x_i$ to its first three nearest neighbors.

Unlike kNN method, no parameter is adopted in NPCM and the connections of each sample may be different from each other. The concept of the SNNs is only used in our method so the proposal is readily understandable. For the datasets with well separation, for example double moons dataset, the cluster results are still dependent on the choice of the parameters in the existed clustering algorithms. However, the good cluster result can be easily obtained by NPCM.

3. Experimental results

In this section, some typical datasets and images are selected to evaluate the performance of the proposed method. These datasets are widely used to test the validity of clustering algorithms. The experimental results show that our proposal can deal with the well-separated datasets with different shape, density and size in case of no parameters. In addition, we can conveniently display the constructed affinity graph by Pajek software.

For the dataset with different density as illustrated in Fig. 1, NPCM can divide it into two clusters properly. This dataset was also tested by various techniques in several algorithms [11,30,31]. It is impossible to correctly find two clusters by FCM and DB scan because of the shape and density. In ref.[11], experimental results indicated that this dataset failed to be clustered by c-means and FCM. It shows that our proposal inherits the merit of kNN method and overcomes the weakness of FCM and DB scan algorithm. Without a priori knowledge about number or type of clusters, our approach can automatically find clusters.
NPCM can also detect the clusters for the datasets with complex shapes and un-equilibrium cluster structure, see for Fig. 2. This dataset introduced by Jain contains 8037 samples and 500 noisy points, and has seven clusters which differ shape, size and density\[11\]. The maximal and minimal size of clusters is C7 (black) with 1600 points and C1 (red) with 299 points, respectively. Two globular clusters on the left side have different density. The well-defined geometric pattern clusters on the right side have cluster center problems. This dataset is a challenge to most clustering algorithms because of these features. It is easily seen from Fig.2 that our method can divide it into seven clusters. This should benefit from more dense connections in the same cluster. Thus, NPCM can handle the datasets with complex cluster structures.
Figure 3. (a) Heterogeneous data set with $|DS|=33$. The density variation in the same cluster is relatively large. Visual inspection shows that three points near the bottom left side of the square may be the outliers. (d) An un-equilibrium data set consists of two great differences in numbers of sample clusters. The large size cluster will be split if we partition it by c-means or FCM method; (g) Three clusters with different density. The density of the right cluster is very different from that of the left two clusters. (b), (c) and (h) The affinity graph constructed by NPCM, respectively; (c), (f) and (i) The corresponding clustering results.

We also test our suggested method on three heterogeneous datasets used in ref. [11,33]. These three data sets were correctly partitioned into groups by some methods ref.[11] (and references therein) if proper parameter values were assigned in advance. Obviously, satisfactory clustering results of these datasets can be obtained by our method without any parameter or threshold. These clustering results demonstrate that the proposed NPCM can
handle the datasets with non-equilibrium structure, or large inter- and intra-class density differences. Tab. 1 lists the cluster numbers obtained by kNN method and our approach. Although the correct numbers of cluster for these three datasets can be found while \( k=3 \) and 4, the problem is how we know these optimal \( k \) in advance?

Table 1. The cluster numbers detected by the proposal and kNN method on three artificial data sets in Fig.3 (ANC--Actual Numbers of Cluster).

<table>
<thead>
<tr>
<th>DS</th>
<th>ANC</th>
<th>3NN</th>
<th>4NN</th>
<th>5NN</th>
<th>6NN</th>
<th>7NN</th>
<th>NPCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig.3(a)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Fig.3(d)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Fig.3(g)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
The clustering results of three real datasets are visualized in Figure 4. The first is a biological example to be relevant groups of genes. This synthetic dataset y14c includes 480 points in 10-d space and has 14 clusters [35]. For this dataset, all the existed clustering algorithms can detect its correct clusters. However the parameters have to be specified previously. The second, Iris dataset [32], is perhaps one of the best known datasets in the pattern recognition. This data set has 150 samples in 4-d space that stand for three physical clusters. For this dataset, two of the three clusters are hardly discernable while the third one is well separated from the other two. Without any class information, the researchers usually argue that there are two, not three clusters in Iris dataset. The third containing 25 grayscale face images is taken from CMU PIE Face Dataset as well as the Yale Face database under different illuminating conditions [36,37]. These examples show that the proposed method can deal with some real classification problem in biology, botany and image classification.
Figure 5. The scatter diagrams displayed in (a), (d) and (g) correspond to the cases N=100, 200 and 500 respectively. The subgraphs shown in (b), (e) and (h) are easily obtained by our approach. These subgraphs exactly reflect their cluster structures exhibited in (c), (f) and (i).

As a standard benchmark, the double moons data set DM(N,d,r,w), is widely used to test the validity of clustering algorithms in pattern recognition, where N is number of samples in each class; d indicates the degree of separation of the two classes (negative value means overlapping); r denotes the radius of ring and w the width of ring. We here take r=10, w=4, and d=-5. Fig. 5 shows the cluster results by applying the proposed approach on double moons. Obviously, the points belonging to one moon should be more similar to each other than those belonging to the other moon. FCM fails to find two correct clusters in DM due to its non-spherical and intertwining structures even if c=2. DBscan method can detect satisfactory clustering result if two parameters, Minps and radius, are set properly.

4. Conclusions and discussions

Based on the shared nearest neighbors, we in this paper propose a nonparametric clustering approach. The suggested method can divide some datasets with different size, shape and density into desired subsets. Our method is easily understandable because only the concept of the shared nearest neighbors is employed. The proposal cannot detect the clusters for the data
sets with overlapped cluster structures or noisy points since it will lead to a connected graph. For the datasets without distinct boundary between two clusters, it is possible to construct a connected graph which can be partitioned further by the algorithms in complex network. But this is out of this article.

NPCM is likely to be more useful in combination with spectral method and dimensional reduction approach, partially due to the more dense links in cluster and the nearest neighbors-preserved. Perhaps the greatest potential lies in applying NPCM to graph-based machine learning algorithms which are widely used in various sciences and engineering fields.

**Conflict of Interests:** The authors declare that there is no conflict of interests regarding the publication of this paper.

**References**


