A Clustering Algorithm Based Absorbing Nearest Neighbors

Hu Jian-jun¹, Tang Chang-jie¹, Peng Jing¹,², Li Chuan¹, Yuan Chang-an¹,³, and Chen An-long¹

¹School of Computer Science, Sichuan University, Chengdu 610064, China
hujianjun@cs.scu.edu.cn, tangchangjie@vip.sina.com
²Department of Science and Technology, Chengdu Public Security Bureau, Chengdu 610017, China
³Department of Information & Technology, Guangxi Teachers Education University, Nanning, Guangxi 530001, China

Abstract. The clustering over various granularities for high dimensional data in arbitrary shape is a challenge in data mining. In this paper Nearest Neighbors Absorbed First (NNAF) clustering algorithm is proposed to solve the problem based on the idea that the objects in the same cluster must be near. The main contribution includes: (1) A theorem of searching nearest neighbors (SNN) is proved. Based on it, SNN algorithms are proposed with time complexity O(n*log(n)) or O(n). They are much faster than the traditional searching nearest neighbors algorithm with O(n²). (2)The clustering algorithm of NNAF to process high dimensional data with arbitrary shape is proposed with time complexity O(n). The experiments show that the new algorithms can process efficiently high dimensional data in arbitrary shape with noisy. They can produce clustering over various granularities quickly with little domain knowledge.

1 Introduction and Background

Clustering Analysis is an important way in knowledge discovery fields. A cluster is a collection of data objects with higher similarity within cluster and lower similarity between clusters. The degree of similarity is usually described by the distance between objects. The greater is the distance, the smaller the similarity is, vice versa. An ideal clustering algorithm should possess scalability, discovery of clusters with arbitrary shape, minimal requirement for input parameters, insensitive to noisy data, ability to deal with high dimension data, interpretability and usability. A lot of clustering algorithms have been proposed in the past years. The existing methods can be classified into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods [1]. The typical algorithms are K-means algorithm [2], CURE algorithm [3], DBSCAN algorithm [4], CLIQUE algorithm [5], BIRCH algorithm [6] [7], etc.. They solve some specific problems by specific methods. However, the multi-layer clustering of high dimensional data with arbitrary shape is also a challenge research fields.

¹ Supported by Grant of National Science Foundation of China (60473071), and Specialized Research Fund for Doctoral Program by the Ministry of Education (SRFDP 20020610007).
This paper proposes an improved nearest distance clustering algorithm—Nearest Neighbors Absorbed First (NNAF) based on the idea of objects in the same cluster must be near.

2 Related Works

The hierarchical clustering algorithms can be classified as being either bottom-up approach or top-down approach. The former starts with each object forming a separate group. Then it successively merges the objects or groups close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The latter begins with all the objects in the same cluster. In each successive interaction, a cluster is split into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds [1]. The typical hierarchical clustering algorithms are BIRCH algorithm, CURE algorithm, Shortest Distance algorithm [8], CHAMALEON algorithm [9], and so on so forth.

There are some common features among NNAF algorithm and Shortest Distance algorithm. In the Shortest Distance algorithm, in order to get the nearest neighbors of an object, you have to compare the distance from it to all other else. Its time complexity is $O (n^2)$. Thus it is not fit for the clustering with large numbers of objects. And it can not process noisy data efficiently. With these statements in mind, NNAF algorithm that can process efficiently clustering with arbitrary shape and noisy data is proposed in this paper. It succeeds in the virtues of Shortest Distance algorithm, which can process high dimensional objects efficiently and nearly require little domain knowledge.

3 Nearest Neighbors Absorbed Firstly (NNAF) Algorithm

3.1 Definitions

**Definition 1.** Given a dataset $V$ in a high dimensional space and distance threshold $d$, where $d > 0$, $V = \{p_1, p_2, \ldots, p_n\}$, then

1. The distance between $p_1$ and $p_2$ is donated as $D(p_1, p_2)$;
2. If $D(p_1, p_2) < D(p_1, p_3) < \cdots < D(p_1, p_n)$, then $p_2$ is the nearest object of $p_1$, namely $p_2$ is the nearest neighbor of $p_1$, which is donated as $\text{NN}(p_1) = p_2$;
3. If $\text{NN}(p_1) = p_2$ and $D(p_1, p_2) \leq d$, then $p_2$ and $p_1$ are of the same cluster.

The basic idea of NNAF algorithm is classifying the two objects satisfying $\text{NN}(p_1) = p_2$ or $\text{NN}(p_2) = p_1$ into the same cluster. Let $\text{NN}(p_1) = p_2$ and $D(p_1, p_2) \leq d$, where $d$ is the distance threshold. If $p_1$ is of the first cluster and $p_2$ belongs to no cluster, then $p_2$ also belong to the first cluster, vice versa. If $p_1$ is of the first cluster and $p_2$ is of the second cluster, then the two clusters recombine into one new cluster and all objects that belong to the two clusters are classified into the new cluster.
Definition 2. Given a dataset \( V = \{p_1, p_2, \ldots, p_n\} \), where \( p_i \in V \), \( p_i \in V \):

1. Each object of \( V \) has a 5-tuple \( \langle \text{Dir}, \text{Distance}, \text{Cluster}, \text{Neighbors}, \text{Reverse-Neighbors} \rangle \), in the terms of \( p_i \), where \( \text{Dir} \) is the coordinates of the point \( p_i \); \( \text{Distance} \) is the distance between \( p_i \) and its nearest neighbor; \( \text{Cluster} \) is the cluster of \( p_i \); \( \text{Neighbors} \) is the set of nearest neighbors of \( p_i \); \( \text{Reverse-Neighbors} \) is the set of objects whose nearest neighbor is \( p_i \).
2. If \( \text{NN}(p_i) = p_k \) and \( \text{NN}(p_k) = p_i \), then \( p_i \) and \( p_k \) are said to be the nearest neighbor to each other.
3. If \( \text{NN}(p_i) = p_k \) and the attributes of cluster of \( p_k \) is empty, then \( p_k \) is said to be the unclassified point of \( p_i \). Otherwise \( p_k \) is the classified point of \( p_i \).

Emerging nearest neighbor to each other or classified nearest neighbor is the clustering termination condition of NNAF.

### 3.2 Nearest Neighbors Searching (SNN) Algorithm

To avoid comparing the distance between one object to all other else when searching its nearest neighbors, our new SNN algorithm tends to find the nearest neighbors fast. Using SNN algorithm, the nearest neighbors can be found by only comparing the distances from one object to the objects in its \( \delta \) domain.

**Theorem 1.** Let \( V = \{p_1, p_2, \ldots, p_{\text{\#}}\} \) be a dataset in \( n \)-dimensional space, where \( p \in V \), \( p_i \) is a object in \( V \). Let \( d \) be the distance threshold. Each dimension of the object with the distance to \( p \) less than \( d \) must be within \( d \) domain in the corresponding dimension of \( p \).

**Corollary 1.** Suppose \( p \) and \( q \) are two objects in \( n \)-dimensional space and the project of the distance from \( p \) to \( q \) on a dimension is \( d \). If \( p \) is in the \( d \) domain region of \( q \), namely \( D(p, q) \leq d \), then the coordinates of \( p \) is same as that of \( q \) on other dimensions.

The space of searching can be reduced dramatically in searching nearest neighbor algorithm based on corollary 1.

**Algorithm 1. Searching Nearest Neighbors (SNN)**

**Input:** Dataset \( V = \{p_1, p_2, \ldots, p_{\text{\#}}\} \), where all objects has been sorted by a certain-dimension, which is denoted by sorted dimension; distance threshold \( d \).

**Output:** The 5-tuple attributes of all objects, where Cluster=NULL, the objects in the attributes of Neighbors and Reverse-Neighbors is that the distance holds the distance threshold.

By SNN, it does not need calculate the distance from one object to all others and only compare the distance from it to the objects in a small domain of it. Thus the computing speed is increased greatly. The time complexity of SNN is \( O(n \log(n)) \). If the objects are obtained by scanning image, the data has been sorted after scanning. So the time complexity will be \( O(n) \). However, that of Shortest Distance algorithm by comparing all distances between every two objects is \( O(n^2) \). After user adjusted the distance threshold \( d \), the algorithm only needs search the nearest neighbors of outliers that have been found in the last time by SNN.
3.3 Nearest Neighbors Absorbed Firstly (NNAF) Algorithm

The idea of NNAF is described as follows: At first, an object Obj1 is classified into a cluster, then all nearest neighbors of Obj1 and the objects whose nearest neighbor is Obj1 are added in the cluster. Finally, all the nearest neighbors of the objects that have been added in the cluster and the objects whose nearest neighbors are the objects that have been added in the cluster are added in the cluster. NNAF iteratively does this process until no objects add in this cluster. In other words, all the nearest neighbors of the objects in this cluster and the objects whose nearest neighbors are the object in this cluster have been added in this cluster. At this time, another unclassified object is classified into a new cluster and performing the clustering follows the above way, until all objects have been added in a certain cluster. The clustering is finished.

Algorithm 3. NNAF algorithm

Input: dataset \( V = \{p_1, p_2, \ldots, p_n\} \), the distance threshold \( d \), the quantity threshold \( q \);

Output: the attribute Cluster of every object

In this algorithm, SNN algorithm is invoked first. Then NNAF algorithm executes following the idea described above.

In NNAF algorithm, getting the clusters need scan the database only once with the time complexity \( O(n) \). And the time complexity of finding the nearest neighbors is \( O(n^{*}\log(n)) \). Therefore the total time complexity is \( O(n^{*}\log(n)) \). When the objects are obtained by scanning image, the time complexity will decrease to \( O(n) \).

3.4 Algorithm Analysis

The algorithm proposed in this paper need little domain knowledge because it only needs two thresholds. The values of the thresholds can be tried many times until the satisfied clustering is obtained.

The clustering based on the nearest neighbors in NNAF has nothing to do with the distributed shape of the objects, the first object selected and the dimensionality. And it needs little domain knowledge and decreases the difficulty for choosing parameters. Therefore, the algorithm fits for clustering with arbitrary shape and high dimensionality. The time complexity of the algorithm is \( O(n) \) since it scans the database only once. The sum of the time complexity is \( O(n^{*}\log(n)) \) after adding the searching nearest neighbors time.

4 Experiments and Results

All of our experiments are conducted on a PC with Intel Pentium III 1GHZ processor and 256 MB memory, which runs Windows XP professional operation system. The initial data of 2-dimensional is shown as Figure 1. It is obviously there are many noisy data and the data distributed with arbitrary shape. It is difficult for many traditional clustering algorithms to process this kind of data. The clustering result is shown as Figure 2 by NNAF.
In this experiment the distance threshold is 2. When it is over 2 the clustering result is nearly the same one. It is because the nearest neighbors can not change however the distance threshold increases when the distance from data to their nearest neighbors are less then the distance threshold. When the distance from most data to their nearest neighbors are more then the distance threshold, the clustering result will be changed dramatically by increasing the distance threshold.

The experimental results show that the algorithms proposed in this paper can solve efficiently the clustering problem with arbitrary shape and noisy data.

Two methods of searching nearest neighbors are used in order to evaluate the performance of SNN. The first one is SNN and the second one is the traditional method. The traditional method is the method which finds the nearest neighbors of every object by comparing all the distances from it to all the other else. The experimental results are shown in Table 1. In Table1 the first column is the number of data, the second and the third columns are the time being used by the two methods respectively. The time unit is millisecond. We can see that SNN is better than the traditional algorithm. It is should be stated that the data comes from scanning Figure 1. The sorting time is saved since the data have been sorted by coordinates after scanning. Although the sorting time is add the SNN is also faster than the traditional algorithm. Because the time complexity of traditional algorithm is $O(n^2)$ and that of the SNN is $O(n \log(n))$.

<table>
<thead>
<tr>
<th>Data Quantity</th>
<th>Time of SNN (ms)</th>
<th>Time of Traditional algorithm (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>39</td>
<td>5587</td>
</tr>
<tr>
<td>4000</td>
<td>69</td>
<td>10825</td>
</tr>
<tr>
<td>6000</td>
<td>100</td>
<td>17314</td>
</tr>
<tr>
<td>8000</td>
<td>129</td>
<td>23052</td>
</tr>
<tr>
<td>10000</td>
<td>160</td>
<td>28892</td>
</tr>
<tr>
<td>10624</td>
<td>179</td>
<td>29161</td>
</tr>
</tbody>
</table>

## 5 Conclusion

In summary, a serial of algorithms about clustering on higher dimension and some related theorems are proposed in this paper. The experiment results indicate that the algorithms are reasonable and can produce satisfied clustering over various granularities. NNAF is also an effective clustering algorithm.
References


