A Spectral Clustering Image Segmentation Algorithm Based on Nyström Approximation

Jian MIAO and Dai CHEN*

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
*Corresponding author

Keywords: Spectral clustering, Nyström approximation, Image segmentation.

Abstract. In the image segmentation, the spectral clustering algorithm are facing the problems, that the similarity matrix and Laplace matrix requires a lot of computing and storage, which limits its expansion in the large-scale image. Aiming at these problems, a semi-supervised spectrum clustering algorithm based on the Nyström approximation was proposed. In the algorithm, adopting the Nyström approximation method to estimate the similarity matrix, which effectively reduces the amount of computation and storage of the spectral clustering algorithm. Finally, on the Brodatz texture library, the image segmentation experiments showed the feasibility and effectiveness of the algorithm.

Introduction

Recently, the image segmentation technology plays an important role in machine learning and artificial intelligence[1]. The spectral clustering method is a clustering algorithm based on dividing the spectrum, which is also very effective in the image segmentation. However, in a large-scale image segmentation, calculating the similarity matrix need very large spending. It will cause problems of high computational complexity and large storage and make the computer memory burdensome.

For the solving problem of similarity matrix of the spectral clustering algorithm in image segmentation, Fowlkes proposed a Nyström approximation method[2]. In recent year, a lot of spectral clustering segmentation algorithms based on the Nyström approximation have been proposed, such as the algorithm in the paper[3-5]. In order to solve the problem that the spectral clustering algorithm has high computational complexity in image segmentation, a spectral clustering algorithm based on the Nyström approximation was proposed.

The organizational structure of the rest of this article is as follows: Section 2 introduces the proposed algorithm, and describes the algorithm steps; Section 3 gives the experimental results, and compares with other algorithms; Section 4 summarizes the work and introduces the future work.

The Spectral Clustering Image Segmentation Algorithm Based on the Nyström Approximation

At present, many image segmentation methods based on spectral clustering mainly focus on the combination of the spectral clustering and other clustering algorithms. Recently, Researchers have proposed a threshold segmentation method based on graph partition which uses the graph partition measures as threshold segmentation criterion to distinguish between target and background[7]. The performance of this method is superior to other types of threshold segmentation method. Such as the Ramesh method [8], the Kapur method[9], etc.

The Basic Idea

In the paper[10], a semi-supervised spectral clustering algorithm based on the Bayesian decision is proposed, that uses the prior information to transform the similarity matrix. Thus, the quality of the clustering results is improved. The idea of the algorithm is to use the Bayesian formula to calculate
the posterior probability $P_{ij}(\omega mlw)$ of each element of the Matrix Dist(the Euclidean distance of data points) belonging to the constraints $M$ and the posterior probability $P_{ij}(\omega clw)$ of each element of the Dist belonging to the constraints $C$. Using the minimum error rate Bayesian decision to divide the element of matrix Dist into Must-link and Cannot-link constraint set.

If the two samples are in $M$ set, the two samples will belong to the same cluster, at this time, using the posterior probability to increase their similarity. On the contrary, if the two samples are in $C$ set, the two samples will belong to the different clusters, at this time, using the posterior probability to decrease their similarity. For a given element $w$, calculate its posterior probability in Must-link and Cannot-link constraint set, and using the posterior probability to give an effective similarity adjusting function. The function need to satisfy the conditions as follows: when the sample satisfies the Must-link constraints, according to its posterior probability belonging to the Must-link constraint set to determine the increase of similarity. When the sample satisfies the Cannot-link constraint set, according to its posterior probability belonging to the Cannot-link constraint set to determine the decrease of similarity. Based on the above conditions, a new similarity adjusting function based on the Bayesian decision is proposed as follows:

$$ W'(x_i, x_j) = \begin{cases} W(x_i, x_j) + P(\omega mlw) - P(\omega clw) \times (1 - W(x_i, x_j)), & P_i(\omega mlw) > P_i(\omega clw) \\ W(x_i, x_j) - P(\omega clw) + P(\omega mlw) \times (1 - W(x_i, x_j)), & P_i(\omega clw) < P_i(\omega mlw) \\ W(x_i, x_j), & \text{otherwise} \end{cases} $$  

A semi-supervised algorithm is clustering on new similarity matrix. Because we can only get a small amount of prior information, and such prior information is only used to adjust the partial similarity between samples, any semi-supervised algorithms cannot ensure clustering results satisfying the prior information constraints. The paper[11] gave two kinds of constraint violations, and proposed a method of adjusting the violating constraint sample subordinate to the cluster.

Given data set $X = \{x_1, x_2, \ldots, x_N\}$ is divided into $k$ clusters, and get $k$ cluster center \{l_1, l_2, \ldots, l_k\}. $y_i = k$ ($k = 1, 2, \ldots, k$) denotes that the cluster label of the sample $x_i$ is $k$. The cluster center of cluster $k$ is marked as $x_k$. In paper[6], the data labels in the clustering result, which violate the constraints are adjusted as follows:

1. The data adjustment of violating the Must-link constraint. Known Samples $(x_i, x_j) \in M$, $l_i \in k$, $l_j \in k'$ ($k$ and $k'$ are two kind of cluster labels.), respectively calculating the sum of similarity between samples $x_i, x_j$ and cluster labels $k, k'$, namely $d_{ik} + d_{ik'}$ and $d_{jk} + d_{jk'}$. If $d_{ik} + d_{ik'} < d_{jk} + d_{jk'}$, the sample $x_j$ will be divided into cluster $k$; Otherwise, the sample $x_i$ will be divided into cluster $k'$.

2. The data adjustment of violating the Cannot-link constraint. Known Samples $(x_i, x_j) \in C$, $l_i \in k$, $l_j \in k$, respectively calculating the similarity between samples $x_i, x_j$ and cluster labels $k, k'$, namely $d_{ik}$ and $d_{jk}$. If $d_{ik} < d_{jk}$, the sample $x_i$ keeps cluster label $k$, and the sample $x_j$ will be divided into the other cluster. Otherwise, the sample $x_j$ will be divided into the other cluster, and the sample $x_j$ keeps cluster label $k$.

This algorithm needs high computing complexity and huge storage space, which limits its application in image segmentation. In order to make this algorithm faster and more efficient, combining a semi-supervised spectral clustering algorithm based on Bayesian decision with the Nyström sampling method, a Bayesian decision semi-supervised spectral clustering algorithm based on the Nyström sampling method(BDN - SSC) is proposed. The main idea of the algorithm is using the Nyström approximation to estimate the similarity matrix to reduce the calculating cost and storage cost of direct calculating similarity matrix.

**Algorithm Process**

Table 1 showed the process of the BDN-SSC algorithm. In the spectral clustering algorithm, the key step is to find the most representative set of eigenvectors. However, the spectrum decomposition time complexity is $O(N^3)$ and space complexity is $O(N^2)$. For the large-scale image, the spectral clustering algorithm is difficult to realize the image segmentation. However, The BDN-SSC algorithm adopts the Nyström approximation method to make the time complexity reduced to...
$O(n^2N)$ and the space complexity reduced to $O(nN)$. Among them, $N$ is the total number of samples, $n$ is the number of randomly selected samples. Obviously, The BDN-SSC algorithm greatly reduced the time complexity and space complexity in image segmentation.

<table>
<thead>
<tr>
<th>Table 1. The BDN-SSC algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong>: The original image, the number of samples $N$, the number of clusters $k$.</td>
</tr>
<tr>
<td><strong>Output</strong>: The image after segmentation.</td>
</tr>
</tbody>
</table>

**Step1**: To deal with the original image and extract the image feature information.  
**Step2**: Extending the prior information sets $M$ and $C$ set, randomly selecting $n$ samples from the image samples. 
**Step3**: Adopting the Nyström approximation method to calculate the similarity matrix $A$ and $B$.  
**Step4**: Using the Bayesian adjusting strategy to adjust the similarity matrix $A$ and $B$. 
**Step5**: Eigenvalue decomposition, chooses the eigenvectors which are corresponding to the first $k$ main eigenvalues $\{v_1,v_2,\ldots,v_k\}$ to structure matrix $V=[v_1,v_2,\ldots,v_k] \in R^{n \times k}$.  
**Step6**: Normalized processing row vectors of the matrix $V$, get matrix $Y$. Each row vector of $Y$ is a sample in the $R^k$ space. Adopting the classic $k$-means to cluster in the space $Y$, and divided $Y$ into $k$ clusters. 
**Step7**: For the samples which violate the constraint information, according to the adjusting strategy dividing the cluster again.

**Experiment**

**The Experimental Data**

The experimental data are the synthetic texture images in the Brodatz repository. The choosing images corresponding category number is 2, 3, 4 and 5. The synthetic texture images are shown in figure 4-1, P1-P5, and the image size is 256×256. The standard segmentation result of images are shown in figure 1, P1-1 to P4-1. In the experiment, the number of samples $N = 200$, and the proportion of prior information is 1%.

**The Evaluation Index**

Generally, given images and the standard segmentation results, the error rate is used as an evaluation index. That the error rate is a simple evaluation index can be more accurate to evaluate the clustering quality. The calculating method is shown in formula (2).

$$\text{error rate} = \frac{\text{error num}}{\text{total num}} \times 100\% = (1 - \frac{\text{accuracy num}}{\text{total num}}) \times 100\%$$ (2)

Among them, the $\text{error num}$ denotes the number of samples which are error classification, the $\text{accuracy num}$ denotes the number of samples which are correct classification, the $\text{total num}$ denotes the total number of samples. The $\text{error rate} \in [0,1]$. The value of $\text{error rate}$ is larger, the more samples are error classification, and the image segmentation results is poorer.

**The Experimental Result and Analysis**

This section is mainly to compare the BDN-SSC algorithm with other two algorithms in the experiment, and analyze the segmenting result. The comparing algorithms are respectively the $k$-means and semi-SSC algorithm. In the figure 1, the P1-2 to P4-2, the P1-3 to P4-3 and the P1-4 to P4-4 are respectively corresponding to the segmentation result of $k$-means, semi-SSC and BDN-SSC algorithm on the texture image.

In the figure 1, the segmentation results of the proposed BDN-SSC algorithm in most cases is superior to the results of $k$-means and semi-SSC algorithm. Although the proposed BDN-SSC algorithm could not avoid the border by mistake, relative to the other two algorithms, the boundary contour is clearer. The results of image segmentation are under the influence of eigenvectors. So the eigenvectors obtained by the proposed BDN-SSC algorithm can better represent the original image structure. In the experimental results, there are still a small amount of discrete points in the comparing algorithm, but the proposed BDN-SSC algorithm can effectively avoid this situation.
The figure 1 respectively showed the texture image segmentation result of the \(k\)-means, semi-SSC and BDN-SSC algorithm. Intuitively, the segmentation results of proposed BDN-SSC algorithm are better than the result of the \(k\)-means and the semi-SSC algorithm. The proposed BDN-SSC algorithm on the segmentation result has good regional consistency. Although segmenting error in the border points, compared with other algorithms, the improving effect of proposed BDN-SSC algorithm is more obvious.

In the experiments, the error rate is the average segmentation result of 5 times. The results are shown in table 2. From the experimental results, the error rate of the BDN-SSC is less than the \(k\)-means and the semi-SSC algorithm on the whole. And through adjusting the class cluster of samples which violates the prior information, the accuracy of the algorithm is much better. The improving effect of proposed BDN-SSC algorithm is great.

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>1.234</td>
<td>1.573</td>
<td>1.334</td>
<td>1.588</td>
</tr>
<tr>
<td>Semi-SSC</td>
<td>0.941</td>
<td>1.364</td>
<td>1.192</td>
<td>1.556</td>
</tr>
<tr>
<td>BDN-SSC</td>
<td>0.862</td>
<td>0.972</td>
<td>1.016</td>
<td>1.482</td>
</tr>
</tbody>
</table>

The proposed BDN-SSC algorithm got the better segmentation result on the texture image, and using a prior information supervision clustering process effectively improved the image segmentation results. Therefore, adopting Nyström approximation strategy is a good way to solve the high computational complexity problem of the spectral clustering in the image segmentation, and improving the quality of clustering is feasible and effective.

**Conclusion**

In this paper, the main work is to study the application of the spectral clustering algorithm in image segmentation. Through adopting the Nyström sampling method, the BD-SSC algorithm is extended into the BDN-SSC algorithm and applied in the image segmentation. In the experiments, compared with the other two segmentation algorithm, the proposed image segmentation algorithm is feasible and effective.
Reference


