

Multi-resolution Hierarchical Clustering by Vector Quantization

Mainak Pal*, Preeti Mukherjee*, Amit Konar

Abstract Clustering aims at grouping of objects or data-points based on a certain measure of similarity. Existing clustering algorithms estimate the measure of similarity of expected data-points to fall in a cluster with respect to presumed or computed cluster centroids. Such approach of distance measure between cluster centroid and possible data points to lie in cluster often results in misclustering, particularly for points equidistant to multiple cluster centroids. This paper offers an interesting solution to this problem by quantization of the attributes of the preferred cluster centroids and then checking the existence of the respective attributes of data-points within the quantized intervals for possible inclusion of data-point in the cluster. This approach, referred to as vector quantization offers additional merits of clustering at different user-defined resolutions of data-points of varying local density. Experiments undertaken confirm the superior performance of the proposed clustering over the state-of-art algorithms with respect to Jaccard coefficient on breast cancer dataset.

1 Introduction

Quantization refers to truncation or round off of an Analog signal to represent it in discrete levels of fixed amplitudes. Clustering, on the other hand, is concerned with grouping of data-points based on certain similarity. A look into standard benchmark dataset reveals that data-points having close proximity in individual attributes, often fall in the same cluster. Unfortunately, the existing clustering algorithms compare the datapoint generally by a distance measure between the cluster centroids and data-points to check the possible existence of data-points within a cluster. Quantiza-

Mainak Pal, Preeti Mukherjee, Amit Konar
Jadavpur University, Kolkata
e-mail: {mainak.pal08, preetimukherjee08}@gmail.com, konaramit@yahoo.co.in

*The two authors contributed equally to this paper.

tion offers the freedom to examine the possible inclusion of individual attributes of a data-point within bounds of the respective attributes of the cluster centroids. This observation inspired the authors to propose a novel algorithm for automatic clustering by quantization of the attributes in the data-points. In this paper, the boundaries of attributes of the data-points in a cluster are selected by a fraction (such as α) of the dynamic range of the respective attributes of all the data-points. The choice of α is left to the user to cluster data at desired resolutions. Thus, the choice of increasing α offers the freedom of multiresolutional hierarchical data-clustering where the lower value of α selects high density clusters (at higher level of the hierarchy) and the larger value of α returns clusters of relatively low data density (at the lower level of hierarchy). In this paper, each data-point is given the freedom to be a trial cluster centroid. However, a few of these data-points, whose (quantized) bounds of attributes include respective attributes of a large number of data-points, concurrently, are selected as the final cluster centroids. Because of the quantized bounds of attributes, two selected cluster centroids always have significant spacing in n -dimensional hyperspace, where n denotes the dimension of the data-point. The proposed algorithm thus facilitates automatic clustering with minimum risk of having false data-points in the clusters. The merit of the proposed algorithm is two fold. First, with the help of quantization of attributes of data-points, it eliminates the chance of false clustering (unexpected data-points) in a given cluster. Second, data-points of different density are clustered at different levels of the hierarchy, thereby providing significant information of data-points within a cluster, obtained at different resolutions. For instance, the high resolution clusters in certain applications, say image-processing, identify objects of interest, whereas the low density clusters offer the inter-relation among objects in the image.

2 Principles and Methodology

2.1 *The proposed clustering Algorithm for vector quantization*

The pseudocode of Vector quantization based clustering is presented below. The pseudocode includes 5 steps. In step 1, the dynamic range of each data dimension/attribute is computed by taking the difference of maximum and minimum values of each attributes in a given set of data-points.

In step 2, each data point is considered as trial cluster centroid, and the possibility that other data-points belong to the cluster of a presumed trial cluster centroid is computed. To accomplish this a neighbourhood of each attribute of a selected data-point is considered and same attribute of other data-points are examined for possible inclusion in the selected range around the selected attribute of the trial cluster centroid. The interval around the attribute of the trial cluster centroid is chosen as “fraction” of the dynamic range of the selected attribute. A set L_k is constructed for the k -th cluster centroid, where the list of point close enough to the cluster centroid

are recorded. Here, by close enough, we mean that the attributes of the data-points must lie in the interval chosen for the respective attribute.

Step 3 is used to set the list of data points in descending order based on cardinality of the set L_k for all k , and top $\eta\%$ cluster of points are selected in descending order of cardinality.

The top $\eta\%$ selected clusters thus obtained may have overlap. Step 4 is used to merge clusters, if there is an overlap of the cluster pairs by more than 90% of either. The process is repeated for all pairs of clusters. The resulting clusters thus obtained are made free from the overlap.

In Step 5, the top ten clusters without overlap are printed.

Algorithm 1 Vector Quantization Based Clustering

Input: A set of l -dimensional data-points $V_k = [V_{k1}, V_{k2}, \dots, V_{kl}]_{1 \leq k \leq m}$ for $k = 1$ to m ; Scale factor $\alpha \in [0, 1]$

Output: Cluster centroids C_1, C_2, \dots, C_n and corresponding data-points in the cluster.

Step 1:

```

1: for each data dimension  $j$  in  $V_k$  do
2:    $DR_j = \max_{1 \leq k \leq m} V_{kj} - \min_{1 \leq k \leq m} V_{kj}$  ▷ Where DR is the dynamic range
3: end for

```

Step 2:

```

1: for each data point  $V_k : k = 1 \rightarrow m$  do
2:   for each cluster  $j$  in  $V_k$  (i.e.  $V_{kj}$ ) do
3:     Initialize  $L_k \leftarrow \phi$ 
4:     if  $V_{k'j} \in V_{kj} \pm \alpha * DR_j$  then ▷ where  $\alpha \in [0, 1]$  is the Scale factor
5:       Then data-point  $k'$  supports data-point  $k$ 
6:        $L_k \leftarrow L_k \cup \{k'\}$  ▷ Save data-point  $k'$  in the set
7:     end if
8:   end for
9: end for

```

Step 3: Sort L_k

```

1: for  $k = 1 \rightarrow m$  do
2:   Sort in descending order of  $|L_k|$ , the cardinality of  $L_k$ 
3:   Take top  $\eta\%$  of the sorted clusters.
4: end for ▷ Rename  $L_k$  as  $L_{\alpha 1}, L_{\alpha 2}, \dots, L_{\alpha r}$  where  $r = (\eta * m) / 100$ 

```

Step 4:

```

1:  $w \leftarrow 1$ 
2: for  $j = 1 \rightarrow r - 1$  do
3:   for  $k = j + 1 \rightarrow r$  do
4:     if  $|L_{\alpha j} \cap L_{\alpha k}| \geq 0.9 * |L_{\alpha k}|$  then
5:       merge  $L_{\alpha j}, L_{\alpha k}$  into  $L_w$ 
6:     else
7:        $L_w \leftarrow L_{\alpha j}$ 
8:        $L_{w+1} \leftarrow L_{\alpha k}$ 
9:     end if
10:     $w \leftarrow w + 1$ 
11:   end for
12: end for

```

Step 5: Print the elements in clusters $L_w \forall w$

2.2 Complexity analysis

In every steps, inner loop operations like addition, subtraction or multiplication with constant predefined entities yields $O(1)$ time complexity.

Now, coming to for loops, as the iterative variables have been incremented by unity, i.e., a constant quantity, so they have $O(n)$ time complexity. Accordingly it can be inferred that step 1 has $O(n)$, while steps 2,4,5 have $O(n^2)$ complexities.

Step 3 is a modified version of Sorting algorithm, which has worst case complexity $O(n \log n)$. Thus, we can conclude that the complexity of our algorithm is $O(n^2)$. Where n denotes the number of datapoints present in the dataset.

3 Multi-resolution Clustering

The clustering algorithm proposed is extended in multi-resolution settings like the well known DBSCAN (Density-based spatial clustering of applications with noise) algorithm. The control in resolution is performed by suitable values of parameter α . The α -value is initialized at 0.2 arbitrarily, and the data-points with high data density are clustered. The resulting clusters obtained for the smallest (possibilities) value of α have highest resolution. After the high resolution data points are clustered, the clustering algorithm is re-invoked again for the second pass with increased value of $\alpha = \alpha + 0.1$. Each time a set of data point of a selected resolution is clustered they are taken out from the list of data-points. The algorithm of multi-resolution terminates, when the number non-clustered data points goes below a user defined threshold. The proposed multi-resolution algorithm is called hierarchical as at different levels of the hierarchy the data-points are clustered at different resolution.

4 Experiment and Results

The prepared Vector-Quantization (VQ) based algorithm has been tested on 32 - dimensional breast cancer dataset for a two class classification problem: malignant and benign. The experiment was repeated for different settings of α in $[0.2, 0.5]$, and it was noted that the proposed VQ clustering out-performs the well known k-means clustering algorithm with respect to Jaccard coefficient.

Results of 32-dimensional data is reduced to 2-dimension using T-SNE (T-distributed Stochastic Neighbor Embedding) technique for visualization and plotted for Ground Truth, KMeans and VQ Clustering for different values of α in Table 1. A table for the Jaccard indices obtained for different values of α for VQ Clustering has been compared with other methods in table 2.

The multi-resolucional hierarchical version of the VQ clustering is tested on synthetic data of sand and sand with white powder, sprayed at different density. It is observed that for $\alpha = 0.2$ to 0.4 all the clusters: pure sand, sand with low density of

white powder and sand with high density of white powder can be clustered easily. Results of this experiment cannot be included here for page restriction. For high dimension, this proposed algorithm outperforms the existing state of art algorithms. The highest Jaccard coefficient has been obtained as **0.83** for $\alpha = 0.24$.

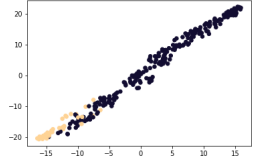
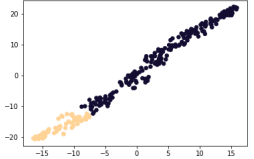
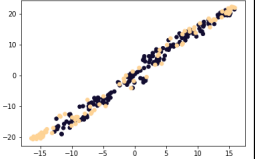
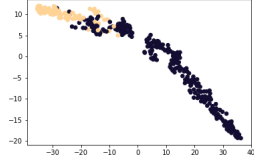
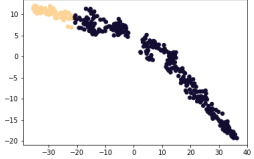
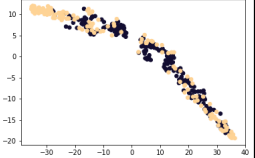
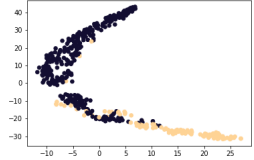
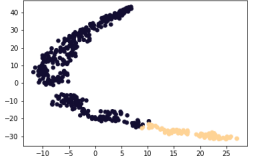
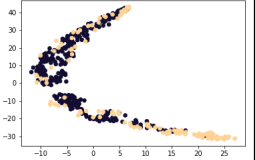
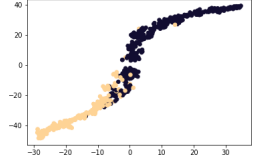
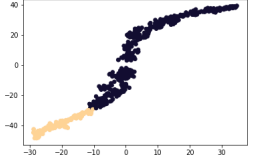
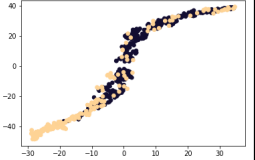
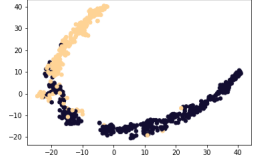
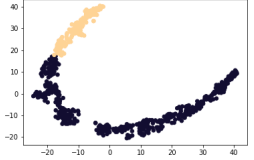
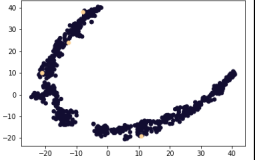
| α | Ground Truth | KMeans | VQC |
|----------|---|---|--|
| 0.20 |  |  Jaccard index = 0.82 |  Jaccard index = 0.72 |
| 0.22 |  |  Jaccard index = 0.82 |  Jaccard index = 0.73 |
| 0.24 |  |  Jaccard index = 0.82 |  Jaccard index = 0.83 |
| 0.26 |  |  Jaccard index = 0.82 |  Jaccard index = 0.82 |
| 0.30 |  |  Jaccard index = 0.82 |  Jaccard index = 0.65 |

Table 1 Jaccard Indices at various scale factors

| α | Jaccard Coefficient | | |
|----------|---------------------|---------------|---------------|
| | K-Means | Fuzzy C-Means | VQ Clustering |
| 0.2 | 0.82 | 0.81 | 0.7200 |
| 0.22 | 0.82 | 0.81 | 0.7300 |
| 0.24 | 0.82 | 0.81 | 0.8297 |
| 0.26 | 0.82 | 0.81 | 0.8176 |
| 0.28 | 0.82 | 0.81 | 0.8127 |
| 0.30 | 0.82 | 0.81 | 0.6490 |

Table 2 Jaccard Indices at different values of scale factor

5 Conclusion

The paper provides a novel and interesting approach of data clustering using vector quantization. Here, the attributes of the data points are quantized into intervals so as to accommodate the respective attributes of other close data-points within the quantized interval of the true cluster centroids. The proposed VQ algorithm has also been extended for multi-resolution hierarchical clustering like DBSCAN. Experiments undertaken on benchmark standard benchmark breast cancer dataset reveals that the proposed technique outperforms the state of art algorithms with respect to standard cluster validation index.

References

1. Wolberg, William H.: Breast Cancer Wisconsin (Diagnostic) Data Set — Diagnostic Wisconsin Breast Cancer Database. In: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
2. R. Radha: Using K-Means Clustering Technique to Study of Breast Cancer. In: R. Radha ; P. Rajendiran. IEEE, World Congress on Computing and Communication Technologies(2014) <https://ieeexplore.ieee.org/document/6755142>
3. Xindong Wu: Top 10 algorithms in data mining In: Xindong Wu,Vipin Kumar,J. Ross Quinlan,Joydeep Ghosh,Qiang Yang,Hiroshi Motoda,Geoffrey J. McLachlan,Angus Ng,Bing Liu,Philip S. Yu,Zhi-Hua Zhou,Michael Steinbach,David J. Hand,Dan Steinberg. Springer-Verlag London Limited (2007) <http://www.cs.uvm.edu/icdm/algorithms/10Algorithms-08.pdf>
4. Martin Ester: A Density-Based Algorithm for Discovering Clusters. In: Martin Ester,Hans-Peter Kriegel,Joerg Sander,Xiaowei Xu. AAAI, KDD-96 Proceedings(1996) <https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>
5. Abdul Hameed: Fuzzy C-Means Algorithm to Diagnose Breast Cancer. In: Dr. W. Abdul Hameed,Dr. Shaik Sharief Basha. International Journal on Recent and Innovation Trends in Computing and Communication