International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-13 Issue-01 May 2024

Finding Number of Clusters before Finding Cluster

Shaik Guntur Mahabub Subhani, Butukuru Rojalakshmi, Vanapamula Veerabrahmachari,

Dr. Godagala Madhava Rao

Associate Professor[1,3], Assistant Professor[2], Professor[4]

subhanimehandi@gmail.com[1],brojalakshmi@gmail.com[2],
vveerabrahmachari@gmail.com[3] ,madhavaraog175@gmail.com[4]

Department of CSE, A.M. Reddy Memorial College of Engineering and Technology, Petlurivaripalem, Narasaraopet, Andhra Pradesh

Abstract –

One of the most pressing issues in cluster analysis is how to count the number of groups in a dataset. Traditional grouping methods usually rely on the user to supply this value. Automatic cluster-number identification is a challenging issue with few good solutions. Some of these methods depend on input from the user, while others make use of computationally intensive cluster validity indices. The visible Assessment of propensity for clustering (VAT) is a relatively new visible method for finding the clustering propensity found in a data collection. We will demonstrate how VAT-based methods can be used for highly effective automated cluster-number determination.

Keywords:

Clustering tendency; genetic algorithm; validity indices; visual assessment.

Introduction

Automatic determination of the number of clusters present in a data set has long been a challenging problem to the researchers. There are two aspects of a clustering problem [1, 2]: finding the number of clusters, and finding the clusters themselves. Majority of the existing clustering techniques assume the number of clusters as an input parameter to be supplied by the user. One of the most common techniques is the k-means algorithm [3]. The k-means algorithm is a crisp partitional clustering algorithm. The objective of crisp clustering is to partition a given data set S containing N data elements { x1, x2, …, xN } into K clusters, C1, C2, …, CK, such that the following conditions are satisfied: Ci □) for all i, Ci ł Cj =) if i □ j, and C1 □ C2 □ … □ CK = S. Besides partitional clustering, there exist hierarchical clustering techniques that can produce a hierarchy of clustering solutions, starting from N number of singleton clusters, having individual data items as the only member of a cluster, up to a single cluster as the complete data set itself. However, in either of the partitional or hierarchical clustering techniques, the unanswered question is which partition or which hierarchy level represents the best clustering solution? This question may be answered if we perform some test for the tendency of clustering of the concerned data set before clustering it.

In many real-life situations the number of clusters in the input data set is not known a priori. Hence, finding out this number before actually applying a clustering algorithm is a real challenge. From commonsense we can say that, for clustering an unlabelled data set should answer the following questions in the given sequence. (1) What is the clustering tendency, i.e., how many clusters (K) are present in the data set? (2) How to partition the data set into K clusters? (3) How to validate the quality of the obtained partition? It is obvious that if K is known a priori, we can apply any known clustering algorithm to obtain a Kpartition. For qualifying a K-partition, many cluster validity metrics (indices) are also available. But when K is unknown, we must solve the first question before solving others. The old and famous ISODATA algorithm [4] uses creation, splitting, merging and deletion of clusters, in repeated steps, to determine an approximate

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-13 Issue-01 May 2024

number of clusters. Each of the above operations depends on some user supplied parameters, about the distribution of data in the data set, which are often very difficult to estimate beforehand. Also, many intermediate clustering solutions are generated and tested during execution of ISODATA – disqualifying it as a candidate to solve question 1. Cluster validity indices are conventionally considered to be useful tools in finding out number of clusters.

These validity indices are generally optimizing in nature, i.e., the optimal value of such a metric (index) identifies the desired number of clusters. But these indices can be applied only after finding out a number of possible partitions. A number of such validity indices exist in literature which include the Dunn's index[5], the DB index [6], the PBM index [7] etc. Unfortunately, validity indices can work only on a pre-computed partition of the data set. It is important to note that finding out a number of possible partitions and then validating them, using a validity measure, is a very time consuming process. But a realistic need of the clustering problem is to determine the number of clusters prior to finding out the corresponding partition, and the process must be fast enough. Speaking differently, we need to assess the actual clustering tendency present in the data set before applying a clustering algorithm. Very recently a visual technique in this regard has been developed [8]. This technique has been referred to as Visual Assessment of Tendency (VAT) for clustering. The VAT process starts with a matrix whose elements are pair wise dissimilarity (distance measures) among the elements of the data set. By reordering the elements of this matrix we get a reordered dissimilarity matrix which tries to accumulate smaller dissimilarity values around the diagonal of the matrix in square contiguous regions. When this reordered dissimilarity matrix is plotted as an image (VAT image), with its elements as pixel intensities, darker square blocks appear along the diagonal line of the image. Each dark block is very closely related with a particular cluster in the original data set. Using this VAT image, we can visually observe the possible number of clusters present in the underlying data set without actually clustering it. However, the visual interpretation part needs to be eliminated from the above procedure, i.e., the process should be fully automated. One major advantage of VAT is that the ordered dissimilarity matrix can be pre-computed, and it has no specific relation with actual partitioning. In this article, we present several methodologies for automatic detection of number of clusters present in a data set which uses, VAT image of the data set as the primary input. These methodologies include visual interactive techniques, image processing based semi-automatic techniques, and GA-based automatic approaches. All these processes have the common goal of automatically finding out the number of dark squares in the VAT image. One important contribution of this article is computation of selected validity indices directly from the VAT dissimilarity data. The rest of the article is organized as follows. In Section 2, we review some of the existing techniques that were in use prior to VAT. Section 3 describes the original VAT algorithm and its use by different techniques. Section 4 provides experimental results.

when two synthetic data sets are used. A comparison of the methods is also provided here. Section 5 contains a conclusion along with some future research direction.

Earlier Works

A number of attempts have been made earlier to estimate the number of clusters present in an unlabeled data set. These methods include split-merge techniques and validity index based techniques mainly. We shall briefly discuss about some of the important techniques before describing VAT based methods. Split-merge based technique: Possibly the oldest attempt in this regard is the ISODATA algorithm [4]. This is a split and merge technique of clustering. Based on number of user supplied information this process tries to develop different possible partitions of a data set by application of split and merge techniques. Although it provides interestingly good results in many situations, its major drawback is the requirement of prior knowledge about the data for determining the externally supplied information. Also, this process actually forms many intermediate partitions during its runtime which makes it a time consuming one, especially for large and complex data sets. Validity index based methods: Here, after formation of a possible partition of a data set, we compute some validity metric for qualifying the partition. Validity indices are generally optimizing in nature, i.e., either the maximum or the minimum value of the metric

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-13 Issue-01 May 2024

represent the best cluster structure (partition). Therefore, these methods consist of finding a number of different partitions of the concerned data set followed by validity computation of each of them. The number of clusters found with the most qualified partition is taken to be the output. Cluster validity indices are generally based on intra cluster compactness and inter cluster separation. They also consider different geometric and statistical properties of the data. Milligan and Cooper [9] provided an elaborate survey of 30 different validity indices, and compared their performances. Some very popular cluster validity indices are the Dunn's Index [5], the Davis-Bouldin (DB) Index [6], and the Pakhira-Bandyopadhyay-Maulik (PBM) Index [7]. Since validity indices can work only after forming a number of competing partitions, these methods are very time consuming and are unable to find number of clusters before clustering. Visual techniques: Visual presentation of clustering information for better understanding of the clustering tendency has been used long ago. These methods include scatter plot of 2-dimensional data, projection of high dimensional data on a plane, converting data dissimilarity values into pixel intensity information and producing dissimilarity image etc. One of the most important contributions toward automation of visual clustering process is due to Ling [10]. In 1973 he has developed a technique called SHADE, which produces a halftone image of hierarchically clustered data (using complete linkage algorithm) for visual display. SHADE may be considered as the first completely automated process for displaying cluster information visually. Many efforts have then been employed for producing a better visual presentation of data dissimilarity. Very recently Bezdek and Hathaway [8] have developed the VAT algorithm (Visual Assessment of Tendency) to display reordered dissimilarity data. A number of variations of the original VAT has been developed which include big-VAT [11], scale-VAT [12], re-VAT [13], and o-VAT [14] algorithms. These variants are mainly supposed to handle vary large data sets efficiently. In the following section we shall discuss the original VAT algorithm in detail.

VAT-Based Algorithms

The Visual Assessment of (Clustering) Tendency (VAT) is a technique for visually analyzing the clustering tendency that is present in the data sets. Different properties and utilities of VAT are discussed in [8]. In case of hierarchical clustering we find another visual technique called SHADE [10] which is a close relative of the VAT algorithm. In visual form VAT data can be displayed as an intensity image.

In VAT we work with a pair wise distance matrix of the original object set O = {o1, o2, …, oN}. In the ijth element of the distance matrix pair wise similarities S = [sij], then dissimilarities can be obtained by a simple transformation, like dij = smax – sij, where, smax denotes the largest similarity value. If the original data set consists of object data X = {x1, x2, …, xN } □ Rm, then dij can be computed as dij = || xi – xj ||, using any convenient norm on Rm. Thus, VAT can be applied over any numerical data set. The original VAT algorithm is presented below. We assume that the dissimilarity matrix D is symmetric having nonnegative off-diagonal entries and zero diagonal entries. In general, the functions, arg max and arg min, in Steps 1 and 2 are set valued, so that the procedure selects any of the optimal arguments. The reordering found by VAT is stored in array P = (P(1), P(2), …, P(N)).

The VAT Algorithm: Input: N × N pair wise dissimilarity matrix D.

Step 1. Set $K=\{1, 2,\ldots, N\}$; Select $(i, j) \in \arg\max\{d_{pq}\}$ where $p,q \in$
Step 2. For $t = 2, 3, \ldots, N$
Select $(i, j) \in \arg\min\{d_{pq}\}$ where $p \in I$, $q \in J$; Set $P(t) = j$
Next t.
Step 3. Form the reordered dissimilarity matrix $\mathbf{R} = [r_{ij}] = [d_{P(i)P(j)}]$,
Step 4. Display **R** as an intensity image, scaled so that $max\ \{r_{ij}\}$ cor

The VAT algorithm rearranges the pair wise distance values in a similar manner to the formation of minimal spanning tree (MST) of a weighted graph following the Prim's algorithm. The difference between VAT and the Prim's algorithm are: VAT does not form a MST, rather it tries to find out the order in which the vertices are added as it is grown, and also it tries to find out the initial vertex which depends on the maximum edge weight in the underlying complete graph. Using this maximum edge weight vertex as the initial point will produce a clear connected graph by avoiding unnecessary zigzagged paths. The permuted indices of the N objects are stored in the array P. Here, no re-computation of distances are done, the reordered graph is obtained by simply rearranging the rows and columns of the original distance matrix. A

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-13 Issue-01 May 2024

sample data set and its corresponding graphical forms and VAT images are shown in Figure 1.

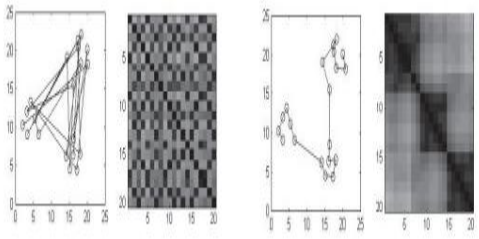

Figure 1. Left: A Sample data displayed as Graph and its dissimilarity image.

Right: Reordered graph and its dissimilarity image. From Figure 1, we can observe the presence of three clusters in the data set which are represented by the three dark square blocks along the diagonal line in Figure 1. The clarity (contrast) of the dark squaredepends on the compactness and separation of clusters in the original data set. In many cases the clusters may be overlapped to some extent. So it is natural that contrast of the VAT image will be lesser in such situations.

Results

We have applied above algorithms over the VAT images of Circular_5_2 and Circular_6_2 data set. These data sets are shown in Figure 2. Circular_5_2 and Circular_6_2 are artificially created 2-

dimensional data sets having 5 and 6 clusters respectively. The former contains 250 elements and the later contains 300 elements. Data are uniformly distributed in both the cases. While for Circular_5_2 clusters are highly overlapped, Circular_6_2 contains relatively disjoint clusters. The data sets and their VAT images are illustrated in Figure 2. We have executed the VGA algorithm for processing VAT images with parameter values: maximum string length = 10, population size = 100, generations = 50, μc = 0.6 and μm = 0.2.

Table 1.Data sets and results

Data Set	Number of clusters obtained with different methods				
	CCE	DBE	VGA with DUNN Index	VGA with DB Index	VGA with PBM Index
Circular_5_2	4	4	2	2	2
Circular_6_2	4	6	4	4	6

Results of experiments with the all above mentioned algorithms are shown in Tables 1. From this table, it is seen that, for the Circular_5_2 data set, all the algorithms fail to determine the number of clusters. This is due to the high degree of overlap present among clusters in this data set. However, 2 or 4 cluster solutions are also very reasonable for this data. For the Circular_6_2 data, either a 4 cluster or a 6 cluster solution is expected. The algorithms provided expected results in all the cases. But, since the CCE and DBE use only image processing techniques, they fail to detect exact solutions. The VGA-based results, for these data sets, show significant improvement. Since, different cluster validity indices have different capacities to resolve among overlapped clusters, different results are obtained. The PBM-index, having good resolution capacity, it gives the best result in this case. It is observed, during experimentation, that time required by the VGA, using the VAT image as its input, is really very small in all the cases. This is a pre-requirement for automatic determination of number of clusters before clustering.

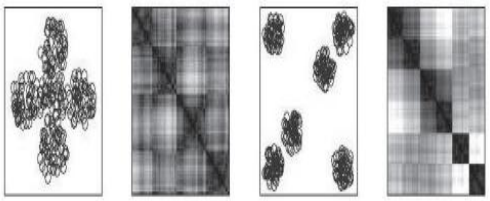

Figure 2. Circular_5_ 2 data , Circular_6_ 2 data, and their VAT images

Conclusions

In this article, we explain a few methods for autonomously determining the number of clusters in a dataset, based on VAT images of the relevant datasets and various known algorithms. When it comes to cluster validity, image processing-based methods fall short because they rely exclusively on the VAT picture structure. Finding the number of groups using traditional validity based methods that

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-13 Issue-01 May 2024

apply directly to the data sets is a time demanding process. The combination of GAs and VAT-based methods is proven to yield effective results rapidly. However, this strategy is dependent on the index's capacity to identify overlapping groups of data. However, this is not a flaw in the GA method itself. This issue can be resolved by employing a reliable authenticity indicator. Additionally, the VAT picture may undergo additional preprocessing, including histogram normalization, grey level lengthening, contrast improvement, noise elimination, etc.

References

[1]. J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles. Reading: Addison-Wesley, 1974.

[2]. A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[3]. J. B. McQueen, "Some methods of classification and analysis in multivariate observations," in Proc. of fifth Barkeley symposium on mathematical and probability, pp. 281-297, 1967.

[4]. G. H. Ball and D. Hall, "Isodata: A novel method of data analysis and pattern classification," tech. rep. Stanford Research Institute, California, 1965.

[5]. J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters," J. Cybern. , Vol. 3, pp. 32-57, 1973.

[6]. D. L. Davis and D. W. Bouldin, "A cluster separation measure", in IEEE Trans. on PAMI, Vol.1, 1979, pp. 224-227.

[7]. M. K. Pakhira, S. Bandyopadhyay and U. Maulik, "Validity Index for Crisp and Fuzzy Clusters," in Pattern Recognition, vol. 37, pp. 487-501, 2004.

[8]. J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in Proc. Intl. Joint Conf. on Neural Networks. Honohulu, HI, pp. 2225-2230, 2002.

[9]. G. W. Milligan and M. C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," in Psychometrika, vo. 50, 985, pp. 159-179.

[10]. R. F. Ling, "A computer generated aid for cluster analysis," in Comm. of ACM, vol. 16, 1973, pp. 335-361. 11. J. C. Bezdek and R. J. Hathaway, "bigVAT: visual assessment of cluster tendency for large data set," in Pattern Recognition, vol. 38, No.

[1]1, pp. 1875-1886, 2005.

[12]. R. Hathaway and J. C. Bezdek and J. M. Huband, "Scalable Visual Assessment of Cluster Tendency," in Pattern Recognition, vol. 39, 2006, pp. 1315-1324.

[13] . J. M. Huband and J. C. Bezdek and R. Hathaway, "Revised Visual assessment of (cluster) tendency (reVAT)," in Proc. of NAFIPS, 2004, pp. 101-104.

[14]. M. K. Pakhira, "Out-of-Core Assessment of Clustering Tendency for Large Data Sets," in Proc. of the nd Int. Conf. on Advance Computing and Communications, 2010, pp. 29-33.

[15]. I. J. Sledge and T. C. Havens and J. M. Huband and J. C. Bezdek and J. M. Keller, "Finding the number of clusters in ordered dissimilarities," in Soft Computing, vol. 13, 2009, pp. 1125-1142.