COMPARISON OF REAL DATASETS CHARACTERISTICS BY USING CLUSTERING APPROACHES

S. Rahamat Basha¹, M.Surya Bhupal Rao², Dr. P. Kiran Kumar Reddy³

¹,² Department of Computer Science Engineering, RGM CET, Nandyal, Andhra Pradesh, India
³ Department of Computer Science Engineering, MRIET, Telangana, India
¹basha.ste@gmail.com, ²suryabhupal@gmail.com, ³kiran.penubaka@gmail.com

Corresponding author: S Rahamat Basha

Abstract

Major issue in cluster analysis is determining the number of clusters present in a data set. The automated identification of the number of clusters can be satisfactorily solved with very few techniques. Recent developments have resulted in a very popular visual mechanism for clustering trend determination (VAT, Visual Assessment of Clustering Tendency) in data sets. The techniques used for image processing depend on the structure of the VAT image, without using any cluster validity concept. High speed solutions can be found in conjunction with GAs from VAT approaches. This approach however depends on the ability of the index concerned to identify overlapping clusters. We will explain how VAT algorithms can be very quickly used to correctly determine the number of clusters. The implementation of the approaches proposed by taking cluster accuracy, cluster error and computational time as metrics.

Keywords: Clustering Analysis, Cluster Accuracy, visual assessment, CCE, DBE, VAT.

I. Introduction

Data mining is the distillation method in large databases with relevant clusters or patterns [VII]. Almost the existing clustering algorithms are inefficient to manage the random dispersion of incredibly large and high dimensioned data sets [I, II, III, VIII] despite the fact that the cluster analysis is a widely-used techniques. The justification is that the statistically based validation methods of cluster analysis acquire a very high computational cost [IX]. The majority of visualization strategies...
are used to make cluster analysis information [XIII, XIV,] instead of analyzing the change in data activity with different algorithm parameters. Due to the fact that visualization is unpredictable, its usefulness in contrasting data grouping is limited.

Clustering algorithms in many disciplines, including healthcare economics, biology marketing and seismic analysis have been successfully applied. Include previously known data comments in a subcategory of clustering algorithms, semi-controlled clustering, used to direct the study of the clustered.

In this work we implement few methodologies in the data set [XV, XVI], VAT images from the data set as the primary input, for automatic detection of the number of clusters. These include visual interaction techniques semi-automatic image recognition and GA-based automatic strategies. All these methods are designed to find the numbers of dark squares in the VAT image automatically [XVII]. The calculation of selected validity indices directly from VAT dissimilarity data will be an important contribution to this article.

II. Related Work

Clustering is a process in which the data is split into various groups of like objects. The objects of a cluster are different from those of other clusters. The number of clusters in data sets is an important and challenging problem. Dataset and clustering processing is a multi-stage process and pre-clustering is important for stressing the relevance of such acts. Clustering steps require fixed information retrieval, for example indexing, scanning, swarming, weighting, etc. The ability of a certain clustering algorithm has to be explored because these phases are necessary for clustering algorithm performance and quality.

As shown below, the clustering system is divided into four levels

1. Data Collection involves clustering, indexing, retrieving and filtering processes. The additional data is also removed.
2. Pre-processing weights and makes data fit for classification and measures data and similarities There are numerous ways to represent the data, for example the graphical model, the data models, etc.
3. The main focus of the study is the pre-clustering process.
4. The main applications used to cluster the data are included in the Post-processing phase. For example, the clustering results allow the recommendation application to suggest news articles for users.
Different pre-clustering algorithms have been proposed for the efficient clustering process, such as DBE, CCE and VAT [I, II, III]. By way of labeling and unlabeling data, the cluster artifacts can be illustrated. In this study, the proposed approaches effectively cluster. The clustering method will lead to different partitioning of a data set based on a particular criteria used. The dataset must be preprocessed before initiating the clustering process. The clustering results and performance of the cluster validity indices from the state-of -the-art pre-clustering algorithms are reported.

Fig. 1: The four stages of the Clustering Process

Fig. 2: Taxonomy of Clustering

Copyright reserved © J. Mech. Cont.& Math. Sci.
S. Rahamat Basha et al
III. Related data on Pre-Clustering

Pre-clustering is a strategy described by Havens et al. (2010). The main point is that error readings (or IRIS error data) should not be sufficient to read correctly (or to correct IRIS data). This is because sequencer errors are unlikely to be replicated by mistake and IRIS results with PCR errors are less amplified. The suggested approach is to assume that H is the correct reading corresponding to R when only a small number of discrepancies have an improved read abundance (H), so that the quantity of R to H is applied and R is discarded. This integration is made with the cluster outs, which at the same time filter chimera and greedy OTSU algorithm[ I, II, III], in order to reduce abundance. To eliminate the extra edges, the pre-processing graph is used. A standard graph clustering approach is used to eliminate the inter-cluster edges. It has been discovered that K-Means clustering reduces noise and shows different results than classical results[XVIII]. Although clustering is studied over a long time period the results of several state of the art algorithms are not satisfactory better algorithms are needed.

When assessing the similarities among the comparable documents, the synonyms of the words used are taken into account but co-occurrence of these words takes priority over synonyms. In certain records, such as news articles. The non-synonymous terms co-occurring in records belong to the same cluster and specifies the degree of co-occurrence in that particular dataset. The method of grouping data into clusters is known as pre-clustering, where cluster objects are similar but different from other cluster objects. As shown in Figure 2, a group of objects is separated into a predefined number of clusters. The dataset is clustered in order to determine the underlying structure and show it as a set of groups. The similarity between subjects must be maximized within a cluster and the resemblance among items from separate clusters should be minimized. Pre-clustering is the grouping by unlabeled data or classified data of similar items. Several methods used in this clustering are post-clustering. Pre-clusters include VAT, DBE, CCE, Co-VAT and Cluster Tendency Method Visual Analysis[I, II, III, IV] as shown in Figure 2-a method suggested by Havens, etcetera (2010)

IV. Visual Techniques Based VAT- Algorithms

CCE algorithm[XI]: The number of clusters from the histogram representation of the generated VAT image is counted for CCE (Cluster Count Extraction). A mixture of various image processing methods is implemented over the VAT image before the histogram is collected. CCE recognizes the reordered N X N similarity matrix R = [rij] and the values of two parameters s and b as inputs. CCE production is the measurement of the number of clusters (equal to the CCE histogram dominant spikes).

The DBEalgorithm [XII]: DBE (Darks Block Extraction), which is processed using several imaging and signal processing techniques, includes the mysterious blocks shown in the VAT file. The DBE accepts the minimum size of a cluster to data size N inputs as an N × N ordered and scaled dissimilarity matrix R = [rij] and a parameter α as inputs.
Since in the dark blocks of the reorganized RDI (VAT image) of the cluster structure, details present in the data set, we must calculate the number of these dark blocks in order for automated cluster number detection. Several methods [XI, X] have been proposed very recently. In [XI] an approach is proposed in which a monotonic transformation enhances the VAT graphic, by comparison. The transformation parameter is the gray level threshold (σ), generated automatically by using Otsu's algorithm over an image of discrepancy.

By using a dissimilarity matrix, however, the VAT algorithm determines the clustering tendency. The diagonal of the VAT image is clearly illustrated with some dark squares. Now we can see that the dark squares represent each cluster. Two types of shapes are found within the VAT image: light rectangles and dark squares. The first one contains wise distance values within the cluster pair, and the second one includes wise distances between the cluster pair. Such distance values can be easily used to determine a validity index of the cluster within and between pairs of clusters. The Dunn's index is one such index of cluster validity. In [XIII] we describe the relevance of the Dunn index and VAT.

V. Results for Numeric and Image dataset

A comprehensive analysis on the validity metrics of clusters for several pre-clustering approaches was performed. Pre-clustering results on the basis of the "Dark blocks" principle for real-time data are created with WINE and IRIS data samples using DBE, CCE and VAT approaches. Compare and select one that best fits the data distribution of the "VAT" for several cluster validation metrics. Cluster validation is the mechanism by which cluster sets are evaluated in terms of their quality and reliability by different pre-clustering processes. Four cluster validation metrics such as clustering accuracy clustering error, time complexity were used in this comparative studies. Table 1. describes two datasets, such as NUMERIC and IMAGE, which consist of features like the classes, attributes, dataset and cluster numbers. Pre-clustering approaches are used in order to detect the number of clusters on various sizes of data sets.

The accuracy of clustering is an easy and straightforward calculation or measurement process. Accuracy is an established consistency assurance criterion. For the purpose of calculating the accuracy, each cluster is determined by counting the number of properly allocated data points and splitting them by N_d to be the class which is the most frequent in that cluster. The accuracy of clustering specifies the clustering quality, as follows:

\[
C_a(N,M) = \frac{1}{N_d} \sum_k \max_j |w_k \cap c_j|
\]

Clustering error determines the relationship between the data points within the cluster. Consider regression results where data is clustered into clusters and regression model errors are independent of clusters, but correlated into clusters. Data point is ‘Rx’ in one cluster, i.e. ‘Re’ is the error correlation within the cluster, and ‘Ca’ is the avg cluster size. If the aggregate regression is present (i.e., Rx= 1), the
The importance of the clustering error is noticed. The following can be used to measure the error of clusters.

\[ Ce = Rx - Re, \text{where } Ce = \text{Clustering Error}, \]
\[ Rx = \text{within cluster correlation regression}, \]
\[ Re = \text{within cluster error correlation regression} \]

Table 1 Datasets and results

| Dataset               | Attributes | Classes | Size(n)       | Number of clusters obtained with various methods |
|-----------------------|------------|---------|---------------|---------------------------------------------------|
|                       |            |         |               | CCE  DBE  VAT                                      |
| NumericDataset N1     | 4          | 6       | 6 * 6         | 4  4  4                                           |
| NumericDataset N2     | 6          | 20      | 20 * 20       | 4  4  4                                           |
| NumericDataset N3     | 6          | 100     | 10000 * 10000 | 6  6  6                                           |
| NumericDataset N4     | 10         | 200     | 20000 * 20000 | 6  6  6                                           |
| ImageDataset I1       | 4          | 6       | 6 * 6         | 4  4  4                                           |
| Image Dataset I2      | 6          | 20      | 20 * 20       | 4  4  4                                           |
| Image Dataset I3      | 6          | 100     | 10000 * 10000 | 6  6  6                                           |
| Image Dataset I4      | 10         | 200     | 20000 * 20000 | 6  6  6                                           |

Table 2 Cluster Approaches Data Sets Numeric, Iris, Wine Vs Cluster Accuracy, Cluster error, Time Complexity

| Cluster Approaches | NUMERIC | IRIS | WINE |
|--------------------|---------|------|------|
| DBE                | CCE     | VAT  | DBE  | CCE  | VAT  | DBE  | CCE  | VAT  |
| Cluster accuracy   | 75.12 % | 60.95 % | 65.53 % | 81.00 % | 70.00 % | 68.00 % | 82.00 % | 72.00 % | 63.00 % |
| Cluster error      | 0.16    | 1.3  | 1.5  | 0.12  | 1.31  | 1.69  | 0.29  | 1.63  | 1.7   |
| Time Complexity    | 3.1 Sec | 4.1 Sec | 3.1 Sec | 4.1 Sec | 4.6 Sec | 3.9 Sec | 4.5 Sec | 3.3 Sec |
Results of experiments with the all above mentioned algorithms are shown in Tables 1. From this table, it is seen that, for the data set, Cluster accuracy, Cluster error, Time Complexity all the to determine the number of clusters. This is due to the high degree of overlap present among clusters in this data set. However, clustersolutions are also very reasonable for this data. For the Table 2, data, either a 4 cluster or a 6 clustersolution 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.

VI. Conclusions

The article describes certain processes in which the numbers of clusters contained in a data set can automatically be classified using various existing algorithms, and VAT data sets can be used as a reference. The techniques used in image-processing depend on the VAT image structure, without using any concept of cluster validity. Conventional techniques based on validity, directly applied on data sets for the determination of number of clusters, take a lot of time. High speed solutions can be found from VAT approaches in conjunction with GAs. However, this method depends on the ability of the index concerned to detect clusters overlapping. But the underlying GA technique is not insufficient. This problem can be alleviated by using a correct validity index. The VAT image could also be fitted with other methods for image processing such as histogram equalization, gray level expansion, increased contrast, noise reduction, etc.

References

I. Ahmad A, Dey L (2007), K-Mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2), 503-527. 2007.

II. Bandyopadhyay S, Saha S (2008), A point symmetry-based clustering technique for automatic evolution of clusters. Knowledge and Data Engineering, IEEE Transactions, 20(11), 1441-1457. 2008.

III. Caliński T, Harabasz J (2012), A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 41(12), 1-27. 2012.

IV. Cattell R (1944) A note on correlation clusters and cluster search methods. Psychometrika, 9(3) (1944) 169-184. 1994.11

V. G.Ravi Kumar, S.Rahamat Basha, Surya Bhupal Rao, “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, 2020.

Copyright reserved © J. Mech. Cont.& Math. Sci.
S. Rahamat Basha et al
VI. 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.

VII. Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek (2009), “Automatically Determining the Number of Clusters in Unlabeled Data Sets”, Fellow, IEEE, 21(3), 335-350.2009.

VIII. L. Wang and C. Leckie and R. Kotagiri and J. C. Bezdek, “Automatically Determining the Number of Clusters in Unlabeled Data Sets,” in IEEE Transaction on Knowledge and Data Systems, vol. 21, 2009, pp. 335-350.

IX. Maimon O, Rokach L (2005), Decomposition methodology for knowledge discovery and data mining: Springer, pp 981-1003, 2005.

X. S. Rahamat Basha, J. Keziya Rani “A Comparative Approach of Dimensionality Reduction Techniques in Text Classification” Engineering, Technology & Applied Science Research, Vol. 9, No. 6, Dec 2019, PP:4974-4979.

XI. S. Rahamat Basha, J. Keziya Rani. JJC Prasad Yadav, “A Novel Summarization-based Approach for Feature Reduction, Enhancing Text Classification Accuracy” Engineering, Technology & Applied Science Research, Vol. 9, No. 6, Dec 2019, PP 5001-5005.

XII. S. Rahamat Basha, J. Keziya Rani, JJC Prasad Yadav, G. Ravi Kumar, “Impact of feature selection techniques in Text Classification: An Experimental study”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-3, September (2019) PP 39-51.

XIII. Surya Bhupal Rao, S. Rahamat Basha, “Chaotic Algorithm for Standard Image Encryption”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-3, September (2019).

XIV. Surya Bhupal Rao, S. Rahamat Basha, G. Ravi Kumar, “A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, 2020.

XV. T. Havens. J. C. Bezdek, J. M. Keller and M. Popescu, “Dunn’s cluster validity index as a contrast measure of VAT images,” in Proc ICPR, Tampa, FL, 2008.

XVI. Timothy C. Havens1, James C. Bezdek1, and James M. Keller1(2012), “A New Implementation of the co-VAT Algorithm for Visual Assessment of Clusters in Rectangular Relational Data”, Fellow, IEEE, 21(3), 335-350.2012.

XVII. Timothy C. Havens, Senior Member (2012), IEEE, and James C. Bezdek, “An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm”, Fellow, IEEE, 21(3), 335-350.2012.

XVIII. Zhang Z, Zhang J, Xue H (2008), Improved K-means clustering algorithm. In Image and Signal Processing, 2008. CISP’08. Congress on, Vol. 5, pp 169-172, 2008.

Copyright reserved © J. Mech. Cont.& Math. Sci.
S. Rahamat Basha et al