AN EFFICIENT PRE CLUSTERING ALGORITHM USING AN UNLABELLED DATA SETS

Uthayakumar Jothilingam 1, Liston Deva Glindis 2

1 P.G. Student, Department of Computer Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, India.
2 Assistant Professor, Department of Computer Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, India.

Abstract: Cluster analysis is one of the primary data analysis methods. The type of Pre clustering algorithm used to estimate the no of clusters in unlabelled data sets. The Selection of the no of clusters is an important and challenging issue in cluster analysis. A no of attempts have been made to estimate no of clusters c in a given data sets. They attempt to choose the best partition from a set of alternative partitions. In contrast tendency assessment attempts to estimate c before clustering occurs. The project focus on pre clustering tendency and determine the no of clusters in unlabelled data sets during cluster analysis by using proposed methodology Trusted Pre cluster Count Algorithm.

Keywords: Cluster Count, Trusted Pre Cluster algorithm, unlabelled data set.

1. INTRODUCTION:
In a data mining community is how to organise a observed data into meaningful structures or taxonomies, considering clustering analysis, it aims at grouping objects of a similar kind into their respective categories

1.1 PRE CLUSTERING TENDENCY ASSESSMENT:
The selection of the no of clusters in an important and challenging issue in cluster analysis. A no of attempts have been made to estimate c in a given data sets.

Most methods are post clustering measures of clusters validity, i.e., they attempt to choose the best partition from a set of alternative partitions.

In contrast tendency assessment attempts to estimate c before clustering occurs. Our focus is on pre clustering tendency assessment bit for completeness, we briefly summarize some existing approached in the post clustering cluster validity problems, describing visual methods for cluster tendency assessment.

To overcome post clustering tendency assessment, Trusted pre clustering count algorithm is introduced for automatically estimating the no of clusters in unlabeled datasets, which is based on the existing algorithms for visual assessment of cluster tendency (VAT) of a data set, using several common image and signal processing techniques such as reVat, Dark block Extraction.

2. EXISTING SYSTEM:
The basic aim of the system analysis is to get the understanding of the needs, what exactly is the need from the software and what are the constraints on the solutions. Analysis leads to the actual specification.

Clustering of unlabelled data poses three major problems.
• Assessing cluster tendency, i.e., how many clusters to seek? or what is the value of c?.
• partitioning the data into c meaningful groups and
• Validating the c clusters discovered.

Many Clustering algorithms require the no of cluster c as an input parameter, so the quality of the resulting clusters is largely dependent on the estimation of c.

To overcome this problem, few existing algorithms are introduced which are facing few drawbacks.

The existing method Dark block Extraction (DBE) is used for automatically estimating the no of clusters in unlabeled data sets, which is based on an existing algorithm for visual assessment of cluster tendency (VAT) of a data set, using several common image and signal processing techniques.

Steps followed for Dark Block Extraction Algorithm.
- Generating a VAT image of an input dissimilarity matrix,
- Performing image segmentation on the VAT image to obtain a binary image, followed by directional morphological filtering.
- Applying a distance transform to the filtered binary image and projecting the pixel values onto the main diagonal axis of the image to form a projection signal, and
- Smoothing the projection signal, computing its first order derivative and then detecting major peaks and valleys in the resulting signal to decide the no of clusters.
2.1 Drawbacks of Existing System

reVAT Methodology
- It becomes hard to mentally integrate the information in a set of c profile graphs when viewed sequentially.
- Clusters in the data are not compact and will be separated, the c profile graphs is pretty confusing.

bigVAT Methodology
- Solves the large data problems suffered by VAT.
- Solves the Interpretation problem solved by reVAT.

Dark Block Extraction Algorithm
- Each and every pixel transformation and calculation occupies a lot of virtual memory space.
- Due to the loss of virtual memory space, it throws error.
- Complexity in where to cut the histogram.

3. PROPOSED SYSTEM:
A new method called TRUSTED PRE CLUSTER COUNT (TPCC) is introduced for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques such as reVAT, bigVAT, Dark Block Extraction Algorithm.

3.1 Steps followed for Trusted Pre Cluster Count Algorithm.
- Generating a VAT image.
- Performing image segmentation on the VAT image.
- Divide the image into matrix pixels as row r X column c.
- Create a proxy controller to store all matrix pixels values and status value.
- Group the similar result's pixels.
- The resulted cluster count will be the perfect pre-cluster count values where it produces the VAT image into a super quality image.

3.2 Advantages of Proposed System:
- TPCC is an advanced method of detecting the number of clusters in a predefined manner in order to give more accuracy to the segmented image.
- TPCC is a pre-clustering method, i.e., it does not require the data to be clustered, nor does it find clusters in the data.
- By using TPCC method, the segmented image is well classified into pixel transformations by maintaining the entire pixel data in a proxy structure, i.e., in an array format. So the calculations process is very little to find the density of the image in order to produce accuracy to the image.

4. FEASIBILITY STUDY:
It is the method to test the system proposal to the workability, impact of the organization, ability to meet user’s needs and effective use of resources.

4.1 Economic Feasibility:
It’s one of the frequently used methods to evaluate the effectiveness of a candidate system. The procedure is to determine the savings and benefits from the candidate system and compare the costs. If the benefits outweigh the costs then it is decided to go ahead with the project.

4.2 Technical Feasibility:
It centers the existing system. It involves financial considerations to accommodate technical enhancements. If the budget is a serious constraint, the project will be judged not feasible.

4.3 Operational Feasibility:
It inherently resistant to change. Computers have been known to facilitate change. It is common knowledge that computer installations have a lot to do with the turnover transfer retaining and changes to employee job status.

4.4 Behavioral Feasibility:
It deals with how the developed software behaves in different scenarios when deployed. It is also a very important part in the different stages of software development.

Input design is a most important part of the overall system design, which requires very careful attention. Often the collection of input data is the most expensive part of the system.

5.1 Segmentation:
5.1.1 Authentication and Authorization:
The authentication is the major part for any kind of software. Generally authentication is used for security purpose to protect from intruders. Here two walls majorly acting for security named as authentication wall and authorization wall.

Authentication wall filters the users by providing username and password with normal rights.
5.1.2 Image Meter:
It acts as a gateway for preview imaging. The main advantage of using the module is to allow scanned machine printed documents and scanned handwritten documents to store.

5.1.3 Preview Imaging:
It will be helpful to the user in order to preview a bunch of scanned images a mingled collection of machine print scanned images and handwritten scanned images stored in the centralized database already.

6. SYSTEM ARCHITECTURE:
It is a conceptual model that defines the structures, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structure of the system which comprises system components.

6.1 Data Flow Diagram:
DFD is a graphical tool used for expressing system requirements in a graphical form. The DFD also known as the bubble chart has the purpose of clarifying system requirements and identifying major transformations that will become programs in system design.

Figure 2: DATA FLOW DIAGRAM

6.1.1 Use Case Diagram:
It’s a Unified Modelling language (UML) is a type of behavioral diagram defined by the created from a use-case analysis.

Figure 3 USE CASE DIAGRAM
7. SYSTEM TESTING:
The main reason behind testing is to find errors. The common view of testing is to bring the program without errors. Software testing is a critical element of a software quality assurance and represents the ultimate review of specification, design and code generation.

7.1 Validation Testing:
It provides the final assurance that software meets all functional behavioral and performance requirements. Validation testing can be defined in many ways, but a simple definition is that validations succeed when the software functions in a manner that is expected by the user.

7.2 Integration Testing:
It is a systematic testing technique for constructing the program structure while at the same time conducting test to uncover errors associated with interfacing. The objective is to take unit – tested modules and build a program structure that has been dictated by design.

8. CONCLUSION:
- Mostly the clusters prefer larger rather than smaller clusters.
- Thus the cluster number extracted by TPCC appears to be increasingly reliable.
- TPCC will probably reach its useful limit when the RDI formed by any recording of D is not from a well-structured dissimilarity matrix.
- Mainly we use Euclidean distance may not be suitable for high dimensional or complex data.
- It is that TPCC does not eliminate the need for cluster validity, but it simply improves the probability of success.
- The initialization of the Trusted Pre Cluster Count Algorithm for object data clustering is highly useful.

9. RESULT ANALYSIS:

Figure 4. VALIDATION TESTING

Figure 5. INTEGRATION TESTING

10. FUTURE IMPLEMENTATION:
- TPCC is more reliable than DBE and CCE.
- The coding has been done more cautiously so that developer can follow the programs easily with the knowledge of the convention followed hence it is easy to be maintained.
- It should not be hard to find an approximate centre sample for each meaningful cluster from any well-structured RDI.
- Inferring the approximate sizes of each cluster.
- It may provide some useful information on object labels, especially for objects around the peak in the projection signals.
- If such label information could be used, only the remaining boundary objects need to be clustered thus the amount of data to be clustered.
11. REFERENCES:


[4] Visualizing class structures of multi-dimensional data by DHIILLION D. MODHA & W SPANGLER.