Computer Science

E-ISSN No : 2581-6195 | Volume : 1 | Issue : 5 | December 2018

AN EFFICIENT PRE CLUSTERING ALGORTHIM USING AN UNLABELLED DATA SETS

Uthayakumar Jothilingam[1], Liston Deva Glindis[2]

ABSTRACT

Cluster analysis is one of the primary data analysis methods the type of Pre clustering algorthim 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 assesement attempts to estimate c before clustering occursThe project focus on pre clustering tendency and determine the no of clusters in unlabeled data sets during cluster analysis by using proposed methodology Trusted Pre cluster Count Algorthim.

KEYWORDS: Cluster Count, Trusted Pre Cluster Algorthim, Unlabelled Data Set

1. INTRODUCTION

In a data mining community is hoe 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 contast tendency assessment attempts to estimate c before clustering occurs.Our focus is on pre clustering tendency asseassment bit for completeness, we briefly summarize some exisiting approched in the post clstering cluster validity problems, describing visual methods for cluster tendency assessment.

To overcome post clustering tendency assessment, Trusted pre clustering countalgorthim is introduced for automatically estimating the no of clusters in unlabeled datasets, which is based on the sxisting algorthims for visual assessmrnt 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 Clusterring algorthims require the no of cluster c as an input parameter, so the quality of the resulting clusters is largrerly dependent on the estimation of c.

To ovecome this problem, few existing algorthims are introduced which are fzacing few drawbacks.

The existing method Dark block Extraction (DBE) is used for aautomatically estimating the no of clusters in unlabeled data sets, which is based on an existing algorthim 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 Algorthim.

- Generating a VAT image of an input dissimilarity matrix,
- Performing image segementation on the vat image to obtainn a binary image, followed by directonal 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 signai, and

[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

HOW TO CITE THIS ARTICLE:
Uthayakumar Jothilingam, Liston Deva Glindis (2018). An Efficient Pre Clustering Algorthim Using An Unlabelled Data Sets, International Educational Journal of Science and Engineering (IEJSE), Vol: 1, Issue: 5, 16-19

Research Paper

- 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 clustes.

2.1 Drawbacks Of Existing System

reVAT Methodolody

- It become hard to mentally integrate the information in a set of c profix graphs when viewed sequently.
- Clusters in the data are not compact and will seperated, 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 Algorthim

- Each and every pixel transformation and calculation occupies lot of virtual memory space,
- Due to 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 teh no of clusters in unlabeled data sets, which is based on an existing algorthim for Visual Assessment of Cluster Tendency (VAT) of a data sets, using several common image and signal processing techniques such as reVAT, bigVAT, Dark Block Extraction Algorthim.

3.1 Steps followed for Trusted Pre Cluster Count Algorthim.

- 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 prox controler 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 no of clusters in a pre definrd manner in order to give more accuracy to the segemented 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 clearly classified into pixel transformations by maintaining the entire pixel data in a proxy structure, i.e., in an array format. So the claculations process is very little to find the density of the image in oret to produce accuracy to the image.

4. FEASIBILITY STUDY

It is the method to test the system proposal to the workability, impact of the organizations, ability to meet user's needs and effective usr of rescources.

4.1 Economic Feasibility

It's one of the frequently used method to evaluate the effectiveness of a andidate 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 ids decided to go ahead with the project.

4.2 Technical Feasibility

It center's the existing system It involves financial considerations to accommodate technical enhancements. If the budget is a serious constraint, the n the project will be judged not feasible.

4.3 Operational Feasibility

It inherently resistant to change the computers have been known to facilitate change. it is common knowledge that com[uter installaions have a lot to do with the turnover transfer retaining and changes to employee job status.

4.4 Behavioral Feasibility

It deals with how to develpe 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. SEGMENTATION

5.1 Authentication And Authorization

The authentication is the major part for any king 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 authorizatin wall.

Authentication wall filters the users by providing username and password with normal rights.

5.2 Image Meter

It acts as a gateway for preview imaging. The main advantage of using the module is to allows yhe scanned machine prinyt document and scanned hand written documents to store.

5.3 Preview Imaging

It will be helpful to the user in order to preview a bunch of scanned images a mingled collection of machinne print scanned images and hand written scanned images stored in the centerallized database already.

6. SYSTEM ARCHITECTURE

It is a conceptual model that defines the sstructures, 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 componets.

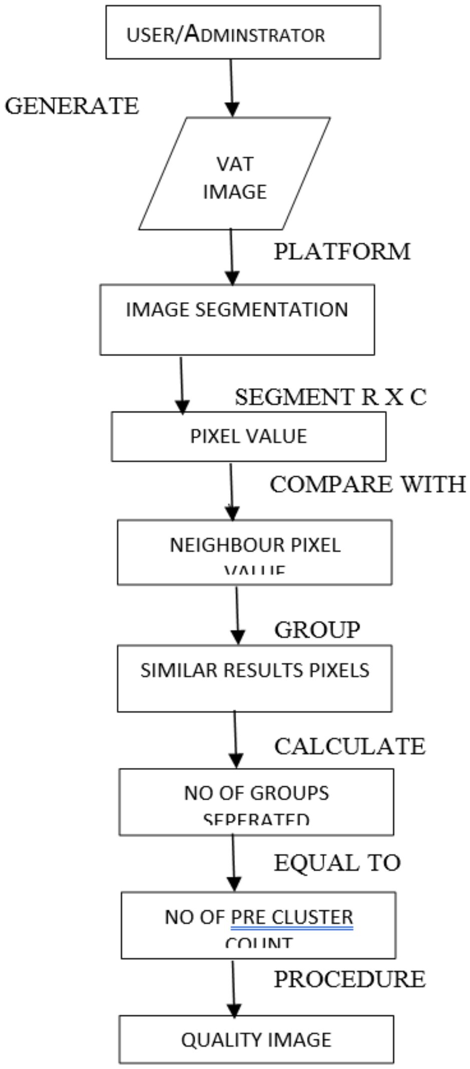


Figure 1: System Architecture

6.1 Data Flow Dagram

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.

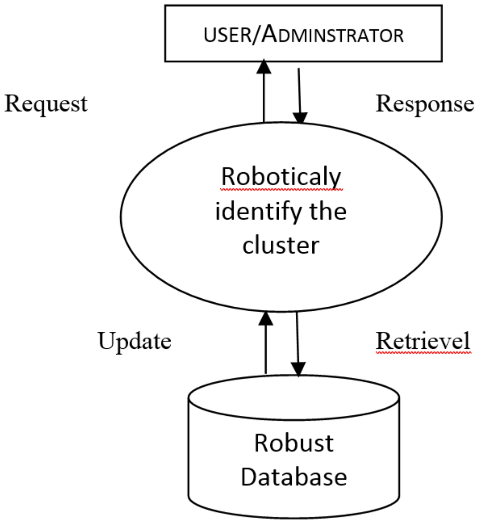


Figure 2: Data Flow Dagram

6.1.1 Use Case Diagram

It's a Inified Modelling language (UML) is a type of behavioral diagram defined by the created from a use-case analysis.

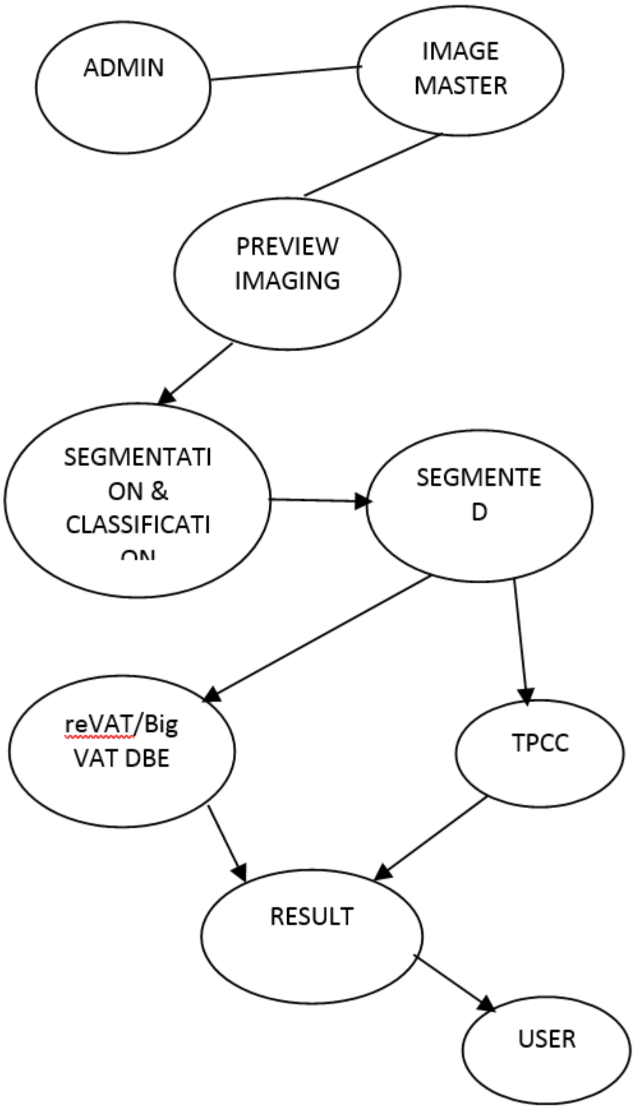


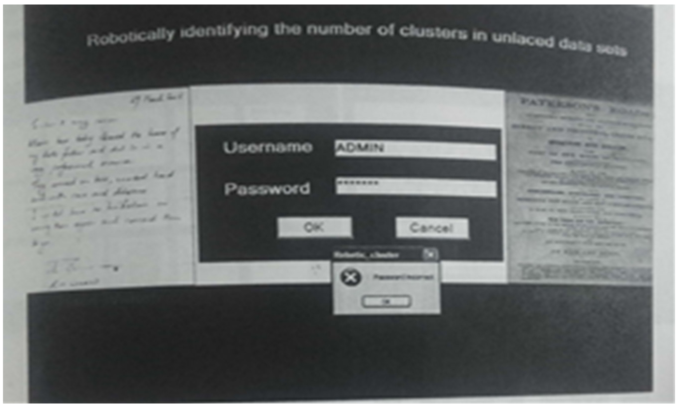
Figure 3: Use Case Diagram

7. SYSTEM TESTING

The main reson behind testind 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 representsthe 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.



Robotically identifying the number of clusters in unlaced data sets

Username ADMIN

Password *******

Figure 4: Validation Testing

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 dicated by design.

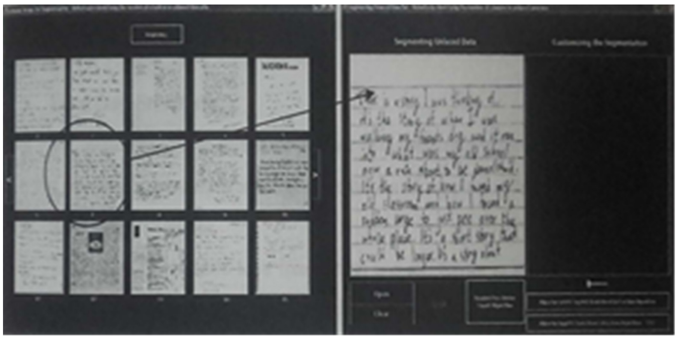


Figure 5: Integration Testing

8. CONCLUSION

- Mostly the the clusters prefer larger rather than smaller clusters.
- Thus the cluster number extracted by TPCC appears to be increasingly reliable.
- TPCC will proprably reach irts useful limit when the RDI formed by any recording of D is not from a well structured dissimilarity martix.
- Mainly we use Euclidean distance may not be suitable for high dimensional or complex data.
- It is that TPCC does not eliminate the nreed fot cluster validity,but it simply improves the probability of sucess.
- The initilization ot the Trusted Pre Cluster Count Algothim for object data clustering is highly useful.

9. RESULT ANALYSIS

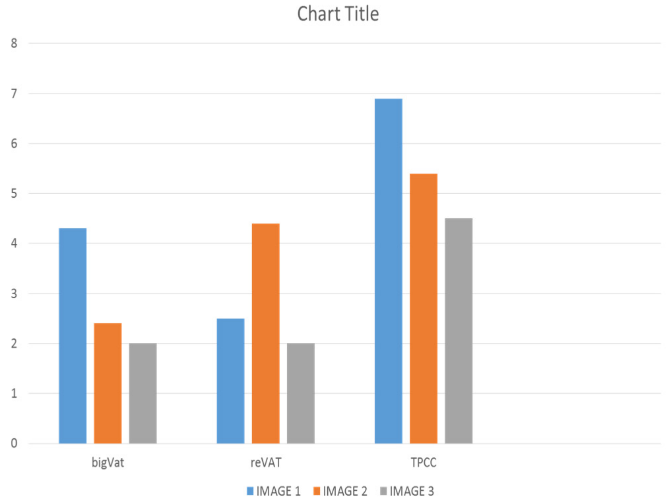


Figure 6: Graph

10. FUTURE IMPLEMENTATION

- TPCC is more reliable than DBE and CCE.
- The coding has been done more cautiously so that developer can follw 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 aproximate 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 objecs need to be clustered thus the amount of data to be clustered.

11. REFERENCES

1. VAT: a tool for visual assessment of (cluster)tendencyieeexplore. ieee.org/document/1007487 by J.C BEZDEK AND R.HATHAWAY.
2. Some new indices of cluster validity IEEE Trans, System,man and Cybernetics by J C BEZDEK and N.R PAL.
3. Geometric approach to cluster validity for normal mixtures by J C BEZDEK and LI. Y ATTIKIOUZEL & M P WINDHAM.
4. Visualizing class structures of multi-dimensional data by DHILLION D , MODHA & W SPANGLER.