Cluster Analysis Research Design model, problems, issues, challenges, trends and tools

V.Ilango[1]
Assistant professor,
Department of Computer Application
New Horizon College of Engineering
Bangalore-103
Email: banalysist@yahoo.com

Dr.R.Subramanian[2]
Professor,
Department of Computer Science
Pondicherry University
Pondicherry

Dr.V.Vasudevan[3]
Professor,
Department of Information Technology
Kalasalingam University
Srivilliputtur-Tamil Nadu

ABSTRACT

Clustering is the process of grouping a set of objects into classes. In the last decade cluster analysis research gained significant interest among researchers. This paper is intended to propose research design model for cluster analysis and to study the problems, various issues that are faced when clustering techniques are implemented .It also considers tools which are readily available and support functions which ease the programming. We also focus on the challenges of clustering analysis and the recent trends for cluster research.

KEY WORDS: Clustering ensembles, Cluster validity, Visualization, Semi-supervised clustering

I INTRODUCTION

Clustering analysis techniques is an interdisciplinary subject. Taxonomists, social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers, and others have all contributed to clustering methodology. The clustering algorithm was driven by biologist Sneath and Sokal in the 1963 in numerical taxonomy before being taken by the stastiscians [30]. Cluster analysis is an iterative process, without any user domain knowledge, it would be inefficient and unintuitive to satisfy specific requirements of application tasks in clustering. The objectives of this paper are to explore and explain the importance and scope of the cluster analysis research. The rest of the paper is organized as follows. In Section 2 clustering analysis research model is discussed. Section 3 describes the major problems, issues and challenges in clustering research. Section 4 and 5 explains recent trends and utilities tools of clustering analysis. The final section concludes the paper.

II CLUSTER ANALYSIS RESEARCH DESIGN MODEL

Research in cluster analysis can express in six-stage model approach. As per Figure.1 the details are discussed. First stage describes the research problem. The primary goal of cluster analysis is to partition a set of objects into two or more groups based on the similarity of the objects for a set of specified characteristics [4]. Second stage explains research design issues, and cluster analysis is also sensitive to outliers. Outliers can represent truly aberrant observation that are not representative of the general population. Similarity is another important concept of cluster analysis. Inter object similarity can be measured in variety of ways, but three methods dominate the application of cluster analysis: correlation measures, distance measures, and association measures. Each of the methods represents a particular perspective on similarity, dependent on both its objectives and type of data. Both the correlation and distance measures require metric data, whereas the association measures are for nonmetric data [5] [6]. In the third stage we must realize that cluster analysis is only as good as the representatives of the sample. Therefore, all efforts should be taken to ensure that the sample is representative and results are generalizable to the population of interest. We must deal with both multicollinearity and

discriminability of the variables to arrive at the best representation of structure. In fourth stage we discuss the clustering algorithm selection methods. The commonly used clustering algorithms can be classified into two general categories: hierarchical and nonhierarchical. Hierarchical procedures involve the construction of a hierarchy of a treelike structure. Nonhierarchical procedures do not involve the tree like construction process [17][25]. Fifth stage is the interpretation stage which involves examination of each cluster in terms of the cluster variate to name or assign a label accurately describing the nature of the clusters. In accessing either correspondence or practical significance, the researcher compares the derived clusters to a preconceived typology. Sixth stage explains validation and profiling of the clusters, validation includes attempts by the researcher to assure that the cluster solution is representative of the general population. In general, the methods of cluster validation are classified into the following three categories: (a) Internal approaches: they assess the clustering results by applying an algorithm with different parameters on a data set and finding the optimal solution,(b) Relative approaches: the idea of relative assessment is based on the evaluation of a clustering structure by comparing it to other clustering schemes and (c) External approaches: the external assessment of clustering is based on the idea that there exists known priori clustered indices produced by a clustering algorithm, and then assessing the consistency of the clustering structures generated by applying the clustering algorithm to different data sets [7][16][27].

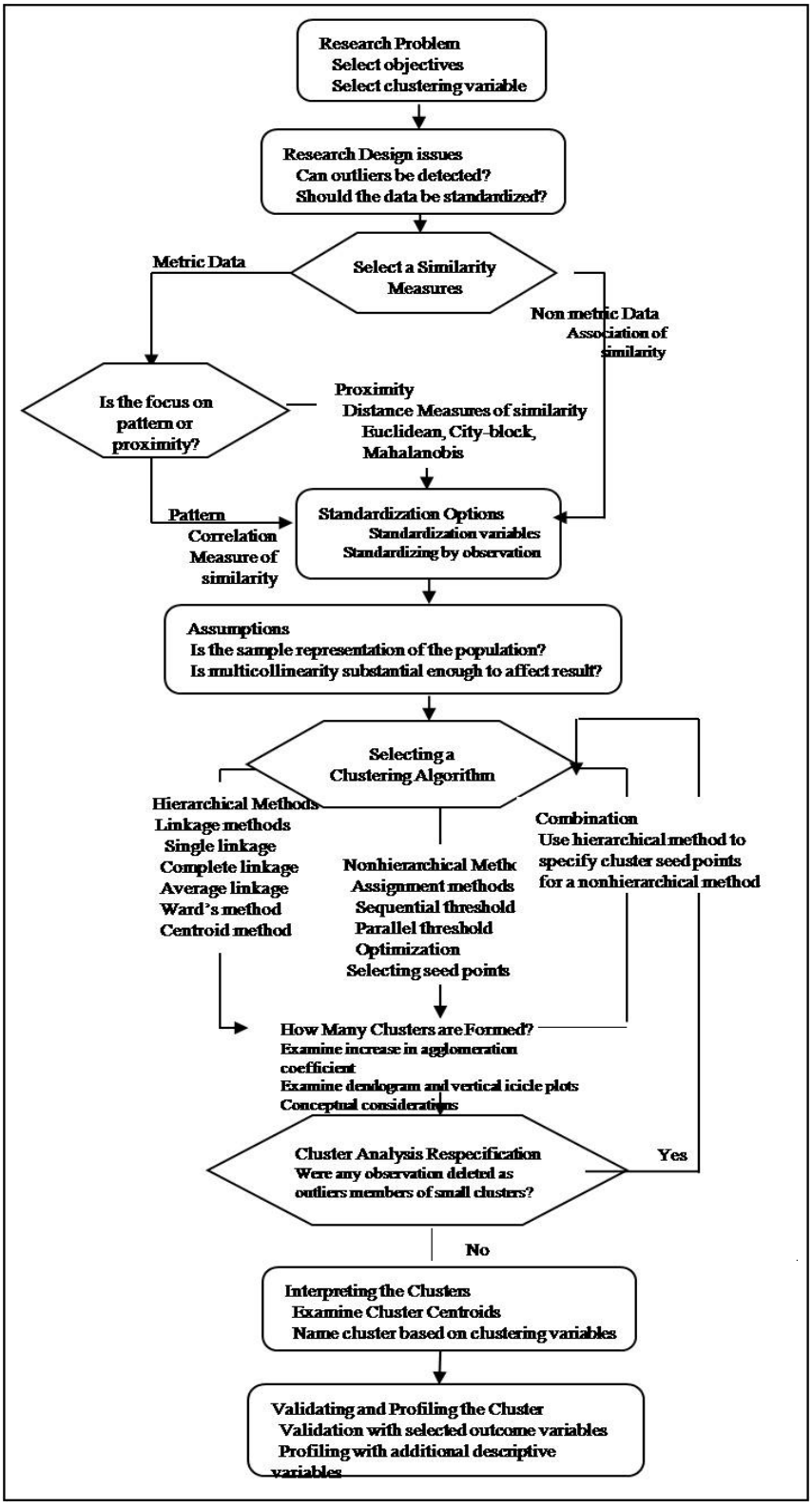

Figure. 1 Clustering Analysis Research Model

III PROBLEMS, ISSUES AND CHALLENGES IN CLUSTER ANALYSIS RESEARCH

A. Problems

The important problems with cluster analysis that we have identified in our survey are as follows:

- *The identification of distance measure :* For numerical attributes, distance measures can b e used. But identification of measure for categorical attributes is difficult.
- *The number of clusters:* Identifying the number of clusters is a difficult task if the number of class labels is not known in advance. A careful analysis of number of clusters is necessary to produce correct results.
- *Structure of database*: Real life data may not always contain clearly identifiable clusters. Also the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect. With a structure less data (for eg. Having lots of missing values), identification of appropriate number of clusters will not yield good results.
- *Types of attributes in a database:* The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.
- *Classification of Clustering Algorithm:* Clustering algorithms can be classified according to the method adopted to define the individual clusters[11][12][13][14][24].

B. Issues in cluster analysis research

The major issues of cluster analysis research can be grouped in the following broad categories [10][18][2][3][22]. Characteristic of data, cluster, clustering algorithm, Data transformation, Cluster solution, Variable selection, Cluster validity and described that subcomponents in table.1

Issues	Description
Data Characteristic	• High dimensionality • Data size • Sparseness • Noise and Outlier • Types of attributes and data set • Scale • Mathematical properties of data space
Cluster Characteristic	• Data distribution • Shape • Differing size • Differing densities • Poorly separated clusters • Relationship among clusters • Subspace clusters
Cluster Algorithm Characteristic	• Order dependence • Nondeterministic • Scalability • Parameter Selection • Transforming the clustering problem to another domain
Data transformation issues	• What measure of similarity/dissimilarity should be used? • Should the data be standardized? • How should non equivalence of metric among variables be addressed? • How should interdependencies in the data be addressed?
Solution issues	• How many clusters should be obtained? • What clustering algorithm should be used? • Should all cases be included in a cluster analysis or should some subset be ignored?
Validity issues	• Is the cluster solution different from what might be expected by chance? • Is the cluster solution reliable or stable across samples? • Are the cluster related to variables other than those used to derive them? Are the clusters useful?
Variable selection issues	• What is the best set of variables for generating a cluster analytic solution?

C .Major Challenges

In existence of large number of clustering algorithms, and their success in a number of different application domains, clustering remains a difficult problem. The following fundamental challenges associated with clustering which are relevant even to this day [8] [9][10][21] are discussed .The researcher must understand the following basic definition before proceeding to research.
(a) What is a cluster?
(b) What features should be used?
(c) Should the data be normalized?
(d) How do we define the pair-wise similarity?
(f) Does the data have any clustering tendency?
(g) Are the all clustering result easily visualized? , Visualization is considered as a collection of transformations from the "problem domain" to the "representation domain". Visualization is the critical challenge of cluster visualization. Cluster analysis should be able to handle several important aspects of visual perception [1]
1. Visualizing large and multidimensional datasets;
2. Providing a clear overview and detailed insight of cluster structure;
3. Having linear time complexity on data mapping from higher dimensional to lower dimensional space;
4. Supporting interactive cluster visual representation dynamically;
5. Involving knowledge of domain experts into the cluster exploration

IV TRENDS AND TOOLS IN DATA CLUSTERING

A. Trends in clustering analysis

Information exploration is not only creating large amounts of data but also a different set of data, like *structured* and *unstructured*. *Unstructured data* is a collection of objects that do not follow a specific format. For example, images, text, audio, video, etc. On the other hand, in *structured data,* there are semantic relationships within each object that are important. A brief summary of some of the recent trends in clustering research on different data sets are briefly discussed below.

1 Clustering ensembles
The success of ensemble methods for supervised learning has motivated the development of ensemble methods for unsupervised learning. The basic idea is that by taking *multiple looks* at the same data, one can generate multiple partitions of the same data [20].

2 Semi-supervised clustering
Any external or *side information* available along with the n x d pattern matrix or the n x n similarity matrix can be extremely useful in finding a good partition of data. Clustering algorithms that utilize such side information are said to be operating in a *semi-supervised mode* . One of the most common methods of specifying the side information is in the form of pair-wise constraints. A *must-link constraint* specifies that the point pair connected by the constraint belong to the same cluster. On the other hand, a *cannot-link constraint* specifies that the point pair connected by the constraint do not belong to the same cluster. [15]

3. Multi-way clustering
Objects or entities to be clustered are often formed by a combination of *related* heterogeneous components. For example, a document is made of words, title, authors, citations, etc. While objects can be converted into a pooled feature vector of its components prior to clustering, it is not a natural representation of the objects and may result in poor clustering performance [23].

4. Heterogeneous data
In traditional pattern recognition settings, a feature vector consists of measurements of different properties of an object. This representation of objects is not a natural representation for several types of data. *Heterogeneous data* refers to the data where the objects may not be *naturally* represented using a fixed length feature vector [26][19].

*Rank Data***:** The task is to rank the cluster data as per properties set by the users whose rankings are similar and also to identify the 'representative rankings' of each group.

Dynamic Data: Dynamic data, as opposed to static data, can change over the course of time e.g., blogs, Web pages, etc. As the data gets modified, clustering must be updated accordingly. A *data stream* is a kind of dynamic data that is transient in nature, and cannot be stored on a disk.

Graph Data: Several objects, such as chemical compounds, protein structures, etc. can be represented most naturally as graphs. Many of the initial efforts in graph clustering have focused on extracting graph features to allow existing clustering algorithms to be applied to the graph feature vectors. The features can be extracted based on patterns such as frequent subgraphs, shortest paths, cycles, and tree-based patterns.

Relational Data: Another area that has attracted considerable interest is clustering relational (network) data.

V .COMMON CLUSTERING TOOLS AND PACKAGES AND PROGRAMME

There are numerous computer software tools both commercial and open source exist. Some of the tools are briefly explained .The list is not limited but many are available [25][28][29].

- BayesiaLab, includes Bayesian classification algorithms for data segmentation and uses Bayesian networks to automatically cluster the variables.
- ClustanGraphics3, hierarchical cluster analysis from the top, with powerful graphics
- CViz Cluster Visualization, for analyzing large high-dimensional datasets; provides full-motion cluster visualization.
- IBM SPSS Modeler, includes Kohonen, Two Step, K-Means clustering algorithms.
- NeuroXL Clusterizer, a fast, powerful and easy-to-use neural network software tool for cluster analysis in Microsoft Excel.
- Neusciences aXi.Kohonen, ActiveX Control for Kohonen Clustering, includes a Delphi interface.
- perSimplex, clustering software based on fuzzy logic. Download available.
- PolyAnalyst, offers clustering based on Localization of Anomalies (LA) algorithm.
- StarProbe, cross-platform, very fast on big data, star schema support, special tools & features for data with rich categorical dimensional information.
- Viscovery explorative data mining modules, with visual cluster analysis, segmentation, and assignment of operational measures to defined segments.
- Visipoint, Self-Organizing Map clustering and visualization.
- Autoclass C, an unsupervised Bayesian classification system from NASA, available for Unix and Windows
- CLUTO, provides a set of partitional clustering algorithms that treat the clustering problem as an optimization process.
- Databionic ESOM Tools, a suite of programs for clustering, visualization, and classification with Emergent Self-Organizing Maps (ESOM).
- David Dowe Mixture Modeling page for modeling statistical distribution by a mixture (or weighted sum) of other distributions.
- MCLUST/EMCLUST, model-based cluster and discriminant analysis, including hierarchical clustering. In Fortran with interface to S-PLUS.
- PermutMatrix, graphical software for clustering and seriation analysis, with several types of hierarchical cluster analysis and several methods to find an optimal reorganization of rows and columns.
- Snob, MML (Minimum Message Length)-based program for clustering
- StarProbe, web-based multi-user server available for academic institutions.
- Weka: Weka is a collection of machine learning algorithms for data mining tasks and is capable of developing new machine learning schemes.
- Matlab Statistical Toolbox It is a collection of tools built on the MATLAB for performing numeric computations.
- Octave: It is a free software similar to Matlab
- SPAETH2-It is a collection of Fortran 90 routines for analysing data by grouping them into clusters
- XLMiner- It has extensive coverage of statistical and machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction.
- DTREG- it is a commercial software for predictive modeling and forecasting offered ,are based on decision trees,SVM,Neural N/W and Gene Expression programs.
- Cluster3-It is an open source clustering software available here contains clustering routines that can be used to analyze gene expression data.
- CLUTO: CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters.

- Clustan: Clustan is an integrated collection of procedures for performing cluster analysis. It helps in designing software for cluster analysis, data mining, market segmentation, and decision trees.

CONCLUSION

In this paper we have covered the research design model for cluster analysis and trends in cluster research. We have also described the problems, challenges and issues faced in implementation and those which affect the clustering results. At last we have described some of the software available that can ease the task of implementation. The study of cluster analysis revealed that there are two major drawbacks that influence the feasibility of cluster analysis in real world applications. The weakness of existing automated clustering techniques lies in dealing with arbitrarily shaped data distribution of the datasets. The evaluation of the quality of clustering results by using statistics-based methods is time consuming when applied on the large database. Therefore several research problems which are not yet adequately solved should be studied in the near future. Special attention should also be given to link the gap between clustering theory and practice of using clustering methods.

REFERENCES

[1] Cheng-Hsuan Li et al, "LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction" , IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 1, FEBRUARY 2011,p.152-163.

[2] Eduardo Raul Hruschka, "A Survey of Evolutionary Algorithms for Clustering", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 39, NO. 2, MARCH 2009, p.123-155.

[3] Fuyuan Cao et al, "A Framework for Clustering Categorical Time-Evolving Data", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 18, NO. 5, OCTOBER 2010,p.872-882

[4] Hair. et.al , " Multivariate data analysis", Pearson Education Pte Ltd, Fifth edition, ISBN:81-297-0021-2

[5] HaiDong Meng et.al, "Research and Implementation of Clustering Algorithm for Arbitrary Clusters", Proceeding of International Conference on Computer Science and Software Engineering, 978-0-7695-3336-0/08,2008.p.255-258

[6] Han and Kamber, " Data mining: concepts and techniques", Morgan Kaufmann publishers, Second edtion,2006, ISBN: 978-1-55860-901-3

[7] H.C. Romesburg. 2004. Cluster Analysis for Researchers. Morrisville, NC: Lulu.com. ISBN 1411606175 / 9781411606173 / 1-4116-0617-5

[8] Hung-Leng Chen, "Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009, p.652-665

[9] Ines Färber et.al , "On Using Class-Labels in Evaluation of Clusterings".ACM -2010

[10] Jain, A.K. Data clustering: 50 years beyond K-means. Pattern Recognition Letter. (2009), doi:10.1016/j.patrec.2009.09.011,p.1-16.

[11] Jie Lian, "A Framework for Evaluating the Performance of Cluster Algorithms for Hierarchical Networks", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 15, NO. 6, DECEMBER 2007,p.1478-1489

[12] Kadim Ta¸sdemir et al, "Topology-Based Hierarchical Clustering of Self-Organizing Maps", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, NO. 3, MARCH 2011,p.474-485.

[13] Madjid Khalilian and Norwati Mustapha, "Data Stream Clustering: Challenges and Issues", Proceeding of International MultiConference of Engineers and Computer Scientists, March 2010, Vol.1

[14] Nancy P. Lin et al, "A Deflected Grid-based Algorithm for Clustering Analysis", INTERNATIONAL JOURNAL of MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, Issue 1, Volume 1, 2007,p.33-39

[15] Nock, R. and Nielsen, F. (2006) "On Weighting Clustering", IEEE Trans. on Pattern Analysis and Machine Inteligence, 28 (8), 1–13.

[16] Osama Abu Abbas, "Comparisons Between Data Clustering Algorithms", The International Arab journal of Information technology, Vol.5, No.3, July 2008, p.320-325

[17] Pang-ning tan et.al , " Introduction to Data mining", Pearson Education, Inc.,2006, ISBN: 978-81-317-1472-0

[18] Rama. B et. Al, "A Survey on clustering Current status and challenging issues", International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2010, 2976-2980

[19] Ren Jingbiao,Yin Shaohong, "Research and Improvement of Clustering Algorithm in Data Mining", Proceeding of *2nd International Conference on Signal Processing Systems (ICSPS), 2010*, 978-1-4244-6893-5,p.842-845

[20] Rui Xu and Donald C. Wunsch, "Clustering Algorithms in Biomedical Research:A Review", IEEE REVIEWS IN BIOMEDICAL ENGINEERING, VOL. 3, 2010,p.120-154

[21] Shi Zhong and Joydeep Ghosh, "A Unified Framework for Model-based Clustering", Journal of Machine Learning Research 4 (2003) 1001-1037

[22] T.Warren Liao, "Clustering of time series data- a survey", Pattern Recognition 38-2005, p.1857-1874

[23] William Claster et.al," Wine Tasting and a Novel Approach to Cluster Analysis", Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, 2010,p.152-157.

[24] Witten and Frank, " Data Mining-Practical Machine learning Tools and Techniques", Morgan Kaufmann Publishers, Second Edition, 2005, ISBN: 0-12-088407-0

[25] http://en.wikipedia.org/wiki/Cluster_analysis

[26] http://www.wikidoc.org/index.php/Data_clustering

[27] http://www.sociosite.net/topics/research.php

[28] http://portal.brint.com/cgi-bin/cgsearch/cgsearch.cgi?query=cluster+analysis

[29] http://www.ploscompbiol.org/article/comments/info%3Adoi%2F10.1371%2Fjournal.pcbi.1001139;

[30] http://en.wikipedia.org/wiki/Numerical_taxonomy