World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:12, 2013

Analysis of Diverse Clustering Tools in Data Mining

S. Sarumathi, N. Shanthi, M. Sharmila

***Abstract*—**Clustering in data mining is an unsupervised learning technique of aggregating the data objects into meaningful groups such that the intra cluster similarity of objects are maximized and inter cluster similarity of objects are minimized. Over the past decades several clustering tools were emerged in which clustering algorithms are inbuilt and are easier to use and extract the expected results. Data mining mainly deals with the huge databases that inflicts on cluster analysis and additional rigorous computational constraints. These challenges pave the way for the emergence of powerful expansive data mining clustering softwares. In this survey, a variety of clustering tools used in data mining are elucidated along with the pros and cons of each software.

***Keywords*—**Cluster Analysis, Clustering Algorithms, Clustering Techniques, Association, Visualization.

I. INTRODUCTION

THE term Clustering is a most fashionable approach and an underpinning task for partitioning a set of data objects into a homogeneous groups or the clusters in Data mining. Clustering method is a useful technique in which it groups the set of data points into clusters such that the points within the single partition are more similar to each other than points in different cluster partition in accordance with the some labeled criteria. The most discrete characteristic of data mining is that it contracts with huge and complex datasets in which its capacity varies from gigabytes to even terabytes. The information in the datasets to be mined often contains the millions of data objects which involves diverse types of attributes or variables such as ratio, binary, ordinal, interval, nominal, categorical etc,. This obligates the data mining operations and algorithms to be robust, stable, and scalable along with the ability to deal different types of the attributes or the objects in the datasets. Hence the process of clustering technique plays a vital role in every aspects of data mining and this leads to the emergence of several clustering tools. From a realistic perspective, the graphical interface used in the clustering tools tends to be more efficient and are more legible to operate which in turn are largely preferred by the clustering practitioners and researchers.

Mrs.S.Sarumathi, Associate Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443321692; e-mail: rishi_saru20@rediffmail.com).

Dr. N. Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

M. Sharmila, PG Scholar, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443581688; e-mail: sharmi28.it@gmail.com).

II. DIFFERENT CLUSTERING TOOLS

A. ClustanGraphics8

ClustanGraphics [1] is an explorative cluster analysis tool for analyzing and solving clustering and also for classification of data problems. Its main aspire is to sort the objects into groups or clusters so that it maintains the degree of association among the members of same cluster. ClustanGraphics provides a most professional task for solving many data analysis process. This software also provides a conventional dendrograms, proximities, scatter plots and profiles. It can offer hierarchical agglomerative cluster analysis and also performs the efficient K-Means algorithm cluster results. ClustanGraphics handles any type of data types even the mixed categories of both numerical and the categorical datasets. Focal point clustering is the new feature involved in this tool which is a two staged K-Means method with random trials.

1. Merits
a) Versatility is the strength of ClustanGraphics
b) Flexibility to handle complex data
c) Simplicity and are ease of use
d) Handling Missing values efficiently
e) Multidimensional Scaling
f) Powerful Visualization of cluster trees
g) No excess baggage requirements needed
2. Limitations
a) Only performs K-Means algorithm
b) Platform dependent
c) Not supported by open source operating systems

B. Cluster 3.0

Cluster 3.0 is a superior version of Cluster which was originally developed at Stanford University. This open source clustering software [2] is mainly designed for the gene data analysis. It provides an efficient graphical user interface for cluster routines accurately. It can also be used directly as GUI program [3] and also can be run in command line prompt. It provides a graphical and computational environment for analyzing data from genomic datasets and also other microarray datasets. The clusters can be organized and visualized in different methods. Initial pre-processing of data can be well equipped in this tool. In this clustering tool most effective tree view of the cluster solutions can be well performed by Java program. The inbuilt cluster programs provides several clustering techniques such as Hierarchical clustering along with its single, complete, average and centroid linkage methods which in turn are highly used for representing the genes in tree form based on their similitude measures. Then well known K-Means algorithm and the Self organizing maps are also builded in this tool to perform their

corresponding tasks of clustering genes into two-dimensional rectangular grids. Finally the evaluation indices are also performed based on the results of the gene clusters.

1. Merits
a) Efficient Pre-processing
b) Flexibility
c) Effective Evaluation measures
d) Ease of Use
e) Platform independent
2. Limitations
a) Only supports Gene datasets
b) Cannot be applicable to other categorical or numerical datasets
c) Large dimensional datasets can't be processed perfectly

C. CViz Cluster Visualization

This CViz Cluster visualization tool [4] is mainly developed and is used for analysing the high dimensional datasets. It can easily load the datasets and the most prominent factor of clusters of the records is displayed effectively. CViz promotes the [5] full motion visualization of the integral data clusters. This tool is highly useful for the researchers who often use the statistical methods. It helps the users in the way of deciding the multiple regression analysis and it is mainly based on the linear discriminate analysis. This software is fully based on Java environment and was developed using it. The Java application itself performs the clustering operations and produces the effective results.

1. Merits
a) Handles both numeric and categorical dataset
b) Platform independent
c) Different types of charts and plots are available
d) Graphs and Animated rotations are the highlighting features
e) Visual images can be displayed
2. Limitations
a) Supports datasets only in CSV format

D. NeuroXL Clusterizer

NeuroXL Clusterizer software [4] is a most affordable, powerful and ultimate ease of use tool for advanced clustering of simple and also complex data. It mainly depends on the highly emerging trend of artificial intelligence and neural network techniques. Through this advancement it delivers a more accurate and stable cluster analysis. This tool is embedded as an add-on to the Microsoft Excel [6] and performs effective clustering on the already existing data in the spreadsheet form. It can also be used extensively in large financial, business and stock market analyst for the task of qualified cluster analysis of the historical data. NeuroXL Clusterizer software [6] also outfits the self organizing neural networks in which it represents the categorization of the trends by recognizing within the dataset given.

1. Merits
a) Easy to use for Beginners
b) No depth knowledge of Neural Networks required
c) Flawless integration with Microsoft Excel
d) Proven Neural Network Technology
e) Maximum Efficiency for affordable cost
2. Limitations
a) Not supported by open source operating systems
b) Large dimension dataset can't be applicable
c) Needs an external user to carry out the classification of variables

E. perSimplex

perSimplex is an intelligent data mining tool [4] mainly used for numeric data processing. The clustering process in this software is mainly based on the Fuzzy logic. From this perspective it can be analysed that it performs artificial intelligence products. The working concept [7] of this perSimplex tool mainly adheres to divide and rule strategy. It brings valuable and reliable information to the user about the nature of the data cluster determined by their various characteristic features. It handles big amount of data and supports the decision making process more effectively. Also performs Data visualization and discovery of data efficiently.

1. Merits
a) Better perception and understand ability of graphical representation
b) Data Processing Speed is very high
c) Ability to handle heterogeneous data
d) User friendly interface
e) Capability to differentiate the similarity levels
f) Having options to control the quality of the outcomes
g) Linear Computing Complexity
2. Limitations
a) Processes only Numerical data
b) Data analytic chart is not suitable for plotting large curves

F. PolyAnalyst

PolyAnalyst is a combination of both statistical pre-processing [4] of data and systematic KDD technique. This software provides the most comprehensible selection of algorithms and techniques for efficient analysis of structured text data. PolyAnalyst mainly works on the certain mechanism [8] in which it builds new functional programs from already existing functional programs. It has a simple drag-and-drop interface which empowers the data analyst and the users to evaluate the data and also performs efficient decision making process. This tool carry out several test mining and analysis tasks [8] such as categorization, taxonomy building, clustering, Natural language search, Entity extraction, Multi-dimensional reporting and Visual link analysis. It undergoes Hierarchical method or Binary grouping of text documents in the task of clustering the data and also promotes automated generation of the tentative taxonomy for categorization. Clustering results can also be evaluated and has the ability to review the solutions using custom changes in the categorization patterns.

1. Merits
a) Text Analysis
b) Efficient Report generation and decision making
c) Discovers hidden values in mass volumes of text data
d) Higher level of scalability

e) Visualization is basically interactive in nature
f) Capability to discover unexpected issues
g) Automated spelling correction
h) Remarkable reduction in the cost of data analysis
i) Good quality and higher speed of data analysis
2. Limitations
a) Only supports text data and are not suitable for numerical data analysis
b) Dictionary editor have to be further enhanced to produce a truly automated system for global analysis
c) Many domain specific terms and definition of words are not included in dictionary manager

G.PermutMatrix

PermutMatrix is a data mining tool [4]-[9] specifically used for analysing and visualizing gene data. It promotes graphical analysis of the dataset and also performs several cluster analysis and techniques for statistical seriation. To analyse and cluster the genes according to the similarities in the gene expression profiles of microarray data, hierarchical clustering methods are highly applied in this tool. In PermutMatrix [10] optimal linear reordering methods such as one-dimensional scaling and seriation, and reorganization of the leaves in the clustering tree are performed effectively. The graphical interface also supports several manual operations as, permutation, inversion, sorting, etc. These processes can be used to locally refine or discover the optimal solutions produced by the above states methods.

1. Merits
a) Interactive graphical interface
b) Simplicity and high efficacy
c) Analyse and produces accurate gene clusters
d) Multiwindow operating environment
e) Large datasets can be quickly analysed
2. Limitations
a) Accepts input datasets only in standard text format and Eisen's cluster format
b) Highly Platform dependent

H.Snob

Snob performs efficient cluster analysis and automatic classification [4] using Minimum Message Length (MML) method, [11] which is a scale invariant Bayesian technique based on information retrieval. It mainly categorizes the datasets based on the numerical distribution and aims to determine the natural classes in the given data. Snob also deals with the finite mixture models such as normal distribution, Poisson distribution, discrete multi-scale distribution, and Von Mises circular distribution. It is highly embedded into WEKA data mining tool and performs efficient clustering results. Snob uses four types of input files to get processes such as Sample files, Variable-Set files, Member files, and Population files.

1. Merits
a) Highly deals and replaces the missing values in datasets
b) It scales and shifts parameters to a great extent
c) Flexibility and are available in different versions supporting C,C++, Java
d) Supports both low dimensional and high dimensional datasets
2. Limitations
a) It only supports numerical datasets
b) Input file formats are highly limited

I. AutoClass C

AutoClass C software [4]-[12] comprises of unsupervised Bayesian classification system that requests for the maximum posterior probability classification. It uses [12] only vector value data in which each instances to be clustered is denoted by the vector of values. These values [13] can be of either real numbers or measurement of attributes. The clusters extracted from the datasets are characterized in terms of the probability distribution over the meta-space defined by each attributes. Real valued attributes are handled through Gaussian distribution model where as Discrete valued attributes are handled through Bernoulli distribution model of AutoClass C. It detects the set of unsupervised classes which are of maximally probable to the data and the model.

1. Merits
a) Concludes the number of clusters automatically
b) Supports both mixed discrete and real valued data
c) Handles the missing values effectively
d) Efficient probabilistic class membership
e) Advanced Report generation based on the cluster outputs
f) Automatic prediction of test case class members from a training classification
2. Limitations
a) AutoClass C is highly limited by memory requirements
b) Excessive Processing time in handling large amounts of data
c) This supports only vector values in the dataset

J. VisiSOM

VisiSOM is a data mining tool [4] for effective clustering and visualization of the multivariate datasets. It mainly aims to detect patterns and classes within the data and creates small subsets of data in order to represent the large dataset. It handles both real and the hypothetical datasets. It includes VisiScreener [14] which is embedded into the software that performs the virtual screening platform. This platform uses the proven Brutus technology within the interactive graphical user interface. The Self Organizing Map (SOM) clustering technique used in this software performs best determination of the shape of clusters and the relative distances between them using the well known unsupervised neural learning algorithm.

1. Merits
a) Explores and deals with complex multi-dimensional data graphically along with 3D view.
b) Replacement of the missing values are effectively done in pre-processing stage
c) Analyses the time dependent data effectively
d) Supports external applications to validate the cluster quality
2. Limitations

a) SOM generate problems if missing values occurs in the dataset
b) Dimensions of the map have to be fixed in advance every time

III. Summarization of Clustering Softwares

The following Table I exemplifies the comparative view of the different clustering tools in data mining based on their compatibility features and its application domain. The main motto of this contrast is not to scrutinize which is the best clustering software but to illustrate the usage paradigm and the awareness of the clustering tools in several areas.

TABLE I
Comparison of Clustering Softwares

Name of the Clustering Software	Platform supported and Installation Prerequisites	Mode of Software	Applications	Types of Dataset supported	Dimensionality of Datasets Supported
ClustanGaphics8	Windows 95, 98, 2000, NT, ME, XP and Mac's via PC Emulator.	Commercial	Statistical Analysis in Social Surveys, Fraud detection and Business domain	Mixed Dataset (both numerical and categorical)	High and low
Cluster3.0	Windows, Mac OS X Linux and Unix	Open Source	Human Genome Centre, and DNA Microarray Analysis	Genomic Dataset	Low
CViz Cluster Visualization	Windows 2000, XP, Vista, 7, 8 and Linux via Java Environment	Open Source	Retail Industries, Web log Sessions and Web document Collections	Mixed Dataset (both numerical and categorical)	High
NeuroXL Clusterizer	Windows 2000, XP, Vista, 7, 8 and Microsoft Excel 2000,2003,2007,2010, 2013.	Commercial	Financial Analysis, Business, Medicine and Research Science	Numerical	Low
perSimplex	Windows XP, Vista, 7, 8 and SimplexDivide.exe, SimplexImpera.exe and SimplexKey.bmp are the necessary files	Commercial	Banking, Insurance, Retail, Energetic, Telecommunication industries and College universities	Numerical	High
PolyAnalyst	Windows XP, Vista, 7, and 8	Commercial	Survey Analysis, Call Center Analysis, Drug safety report Analysis, Incident Report Analysis, market research to bioinformatics research and development of physical models	Text Dataset	High
PermutMatrix	Windows 2000, XP, Vista, 7, 8 and MS Visual C++ needs to be installed	Open Source	Bioinformatics industries and DNA microarray analysis	Genomic Dataset	High and Low
Snob	Windows XP, Unix and Linux	Open Source	Financial and Retail industries, Business profit Analysis	Numerical	High and Low
AutoClass C	Windows 95, 98, 2000, NT, XP, Unix and Mac OSX platforms	Open Source	Intelligent System Division, military, and Business industries	Mixed Dataset (both discrete and real valued data)	High and Low
VisiSOM	Windows 2000, XP, Vista, 7, 8 and Linux , 32-bit x86 Compatible CPU along with an Ethernet Card	Commercial	Pharmaceutical Industries, Biochemistry analysis and Environmental surveys	Multi-dimensional Dataset	High and Low

IV. Conclusion

In this paper, various clustering softwares available for the purpose of several tasks in data mining are highly elucidated. Clustering technique tends to be the most underpinning process in each and every phase of information retrieval under various domains. The immense significance of clustering process leads to the innovation in the field of software development. Clustering softwares has the extensive technical paradigm and inbuilt complex algorithms which is very useful for handling the massive amount of data more precisely and efficiently. Additionally the graphical interface in the clustering tool outperforms the traditional clustering program through its user friendly nature, processing speed, and the most professional way of representing the cluster results. Thus the main contribution of this paper is to promote knowledge about the different clustering softwares and its applications in various industries which will be very useful for the readers and also it meets the needs of researchers to innovate more clustering tools in future.

References

[1] Clustan Graphics Homepage [Online]. Available at: http://www.clustan.com/clustangraphics.html/
[2] Open Source Clustering software [Online]. Available: http://bonsai.hgc.jp/mdehoon/software/cluster/software.html/
[3] Michael Eisen and Michiel de Hoon, "Cluster 3.0 Manual" University of Tokyo, Human Genome Center, 2002.
[4] Software: Clustering and Segmentation of data mining and analytics [Online].Available: http://www.kdnuggets.com/software/clustering.html/
[5] Data mining-CViz Software-Data Mine Wiki [Online]. Available: http://www.the-data-mine.com/CvizSoftware/
[6] Neuro XL Clusterizer Software [Online]. Available: http://neuroxl.com/products/excel-cluster-analysis-software/.
[7] perSimplex[Online].Available: http://www.persimplex.biz/index.php
[8] Mikhail V. Kiselev "PolyAnalyst 2.0: Combination of Statistical Data pre-processing and Symbolic KDD Technique" Proceedings of ECML-95 Workshop on Statistics, Machine learning and Knowledge discovery in Databases. Heraklion Greece, pp 187-192, 1995.
[9] PermutMatrix Software [Online]. Available: http://www.atgc-montpellier.fr/permutmatrix/
[10] Gilles Caraux and Sylvie Pinloche "PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order", Bioinformatics Application Note, Vol.21 No. 7 pages 1280-1281 doi number: 10.1093/bioinformatics/bti141, 2005.

[11] Snob Software Homepage [Online]. Available: http://www.csse.monash.edu.au/~dld/Snob.html
[12] Data mining-Auto Class C-Data Mine Wiki [Online]. Available: http://www.the-data-mine.com/AutoClass C/
[13] AutoClassC Software Homepage [Online]. Available: http://ti.arc.nasa.gov/tech/rse/synthesis-projects/autoclass/autoclass-c/
[14] VisiSOM- Visipoint Software homepage [Online]. Available: http://www.visipoint.fi/visisom.php

Mrs. S.Sarumathi received B.E. degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E. degree in Computer Science and Engineering from K.S.Rangasamy College of Technology, Namakkal Tamil Nadu, India in 2007. She is doing her Ph.D. programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 15 years. At present she is working as Associate professor in Information Technology department at K.S.Rangasamy College of technology. She has published 2 reputed International Journal and two National journals. And also she has presented papers in three International conferences and four national Conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Dr.N.Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in 2001. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor & Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 14 papers in the reputed international journals and 9 papers in the national and international conferences. She has published 2 books. She is supervising 14 research scholars under Anna University, Chennai. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, Pattern Recognition and Network security. She is a life member of ISTE.

Miss. M.Sharmila holds a B.Tech degree in Information Technology from K.S.Rangasamy College of technology, affiliated to Anna University of Technology Coimbatore, Tamil Nadu, India in 2012. Now she is an M.Tech student of Information Technology department in K.S.Rangasamy College of Technology. She has published 1 international journal and presented three papers in National level technical symposium. She is an active member of ISTE. Her Research interests include Mining Medical data, Opinion Mining and Web mining. Most of her current work involves the development of efficient cluster ensemble algorithms for extracting accurate clusters in large dimensional database.