iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** VI **Month of publication:** June 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

An Insight Into Concept ofCluster Analysis and Its Evaluation.

Nidhi Chawla[1] and Ajmer Singh[2]

[1]M.Tech Research Scholar, Department of Computer Science and Engineering
DeenbandhuChhotu Ram University of Science and Technology,Murthal, Sonepat

[2]Assistant Professor, Department of Computer Science and Engineering
DeenbandhuChhotu Ram University of Science and Technology, Murthal, Sonepat

Abstract -Cluster analysis is a method for identification of homogeneous groupsof objects (observations, cases or events) based on the information found in the data describing the objects or their relationships. Themain goal is to find group of objects such that objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. If the similarity (or homogeneity) within a group and the difference between groups will be greater, there will be "better" or more distinct clustering. One of the clustering process issues is the evaluation of clustering results. Estimation of the obtained cluster structure quality is the main subject of cluster validity. This paper introduces the fundamental concepts of clustering and addresses an important issue of clustering process. This paper also describes various clustering methods, types of clusters and validity index used by clustering to categorize complex data into real and accurate knowledge.Furthermore it illustrates the variousIssues , Problems challenges and Tools of clustering analysis.

Keywords :-Clustering, Types of Cluster, Validation Indexes, Challenges.

1.INTRODUCTION

Clustering is the useful tasks in data mining process because it play a key role for analysis and exploration of data. Cluster analysis techniques is an active topic of research and its applications are found in many fields including pattern recognition, image processing, spatial data analysis, market research ,www, and machine learning[1].Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic .The process of clustering is also known as un supervisory or machine learning .The objectives of this paper are to explore and explain the importance and scope of the cluster analysis . The rest of the paper is organized as follows. In Section 2 wehave given brief overview of different type of clusters and Clustering methods .Section 3 describes the major problems, issues and challenges in clusteringanalysis. Section 4 has several validation index for testing tools of Clustering Algorithms in Section 5 and Conclusion in Section 6.

2. CLUSTER ANALYSIS

Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into

ISSN: 2321-9653

groups (clusters). Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction[1][2].

2.1) Clustering Process

Clustering process is mainly used for revealing the "sensible groups" from organization of patterns in order to discover similarities and differences, as well as to evaluate useful conclusions about them. Clustering process is known as an unsupervised process because there are no predefined classes, labels and examples that would show what kind of desirable relations should be valid among the data.Thus, there is a need of preprocessing before we assume a clustering task in a data set[10]. So there are basic steps of clustering process which is presented in figure 1 and can be summerised as follow[3]

1)Feature selection. The goal is toselect subset of input variables by eliminating those features which has little or no predictive information.Feature selection is required due to presence of noisy, irrelevant or misleading features. Thus, preprocessing of data may be necessary prior to theirutilization in clustering task. Preprocessing involves transformation, discretization and normalization of data.

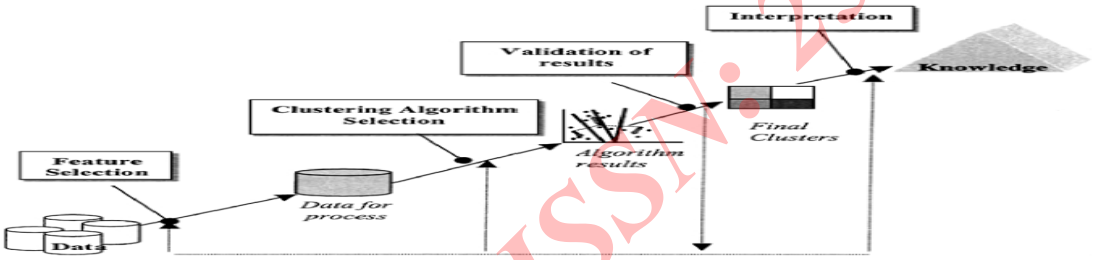


Figure 1. Steps of clustering process.

2)Clustering algorithm. This step refers to the selection of an algorithm which results in the definition of a good clustering scheme for a data set. This step is the combination of selection of corresponding proximity measure and construction of a clustering criterion function. .

i)Proximity measure: Itis a measure that determinethe "similarity" of two data points (i.e. feature vectors) .Proximity measure directly effects the formation of cluster. In most of the cases we must ensure that all selected features equally contribute to the computation of the proximity measure and no features will dominate or effect others.

ii)Clustering criterion: In this step, clustering criterion is defined, which can be expressed via a cost function or some other type of rules. Thus, It may be defined that a "good" clustering criterion, lead to a partitioning of clusters that fits well for the data set.

3)Validation of the results. The goodness and correctness of clustering algorithm results is verified by using appropriate criteria and techniques. Since in clustering algorithms , clusters are not known a priori, irrespective of the clustering methods, t some kind of evaluation is required for final partition of data in most of the application

4)Interpretation of the results.The ultimate goal of clustering is to extract meaningful knowledge from the original data.In most of cases, the experts in the application area integrate the clustering results with other experimental evidence in order to draw the right conclusion.

2.2)Different Types of Clustering Method

ISSN: 2321-9653

A multitude of clustering methods are proposed in the literature. Clustering algorithms can be classified according to:

- The type of data which is used for the algorithm for input.
- The criteriaof clusteringwhich define the similarity between data points.
- The theoretical and fundamental concepts on which techniquesof clusteranalysis are based (e.g. fuzzy theory, statistics).

Clustering algorithms have been applied to a range of problems in a wide variety of research fields.Thus according to the method adopted to define clusters, the algorithms can be broadly classified into the following types [3][4][9][11][12]

1) Partitional clustering attempts to directly partition the data set into a set of disjoint clusters. More specifically, they attempt to determine an integer number of partitions that optimise a certain criterion function. The local or global structure of the data is emphasized bythe criterion function andoptimization of cluster is an iterative procedure. eg are K Means and K-Medoids algorithm

2) Hierarchical clustering proceeds by either merging the smaller clusters into larger ones, or by splitting larger clusters into smaller ones. The result of the algorithm is a tree of clusters which is called dendrogram.Itshowsthat how clusters are related to each other., a clustering of the data items into disjoint groups is obtained by cutting the dendrogram at a desired level.eg are BIRCH and CURE algorithm

figure 2: Three well-separated clusters of 2 dimensional points.

2) Center-based Cluster :A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid as described in figure 3.

3) Density-based clustering. In this type of clustering there is grouping of neighbour objects of a data set into arbitrary shaped clusters based on density conditions.eg are DBSCAN,OPTICS and DENCLUE algorithm

4) Grid-basedclustering.This type of algorithms is mainly proposed for spatial determining because it uses a multiresolution grid data structure . Their main characteristic is that the space is quantised into a finite number of cells and then all operations is done on the quantised space. eg are STING and CLIQUE algorithm.

2.3) Different Types of Clusters

The main objectives of clustering method for a data set is to discover existing significant groups. These methods usually look for the clusters which contain objects as closer to each other as possible and this indicates a high level of objects similarity.[1] .The definition of what constitutes a cluster is not well defined, However, several working definitions of a cluster are commonly used.

1)Well-Separated Cluster :A cluster is set of objects in which each objectin a cluster is closer (or more similar) to every other object or pointin the cluster than to any objectnot in the cluster. Well separated cluster can be of any shape, need not to be of globular shape as described in figure 2.

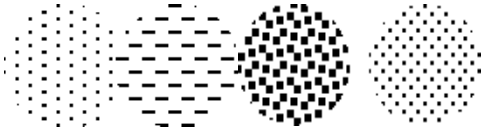


figure 3: Four center-based clusters of 2 dimensional points

3)Contiguous Cluster (Nearest neighbor or Transitive Clustering): A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points

ISSN: 2321-9653

in the cluster than to any point not in the cluster as described in figure 4

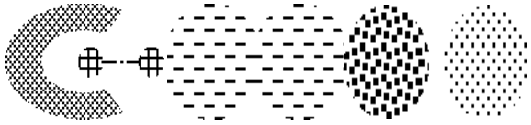


Figure 4: Eight contiguous clusters of 2 dimensional points.

4) Density-basedCluster:A cluster is a dense region of ints, which is separated by low-density regions, from other regions of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present as described in figure 5.

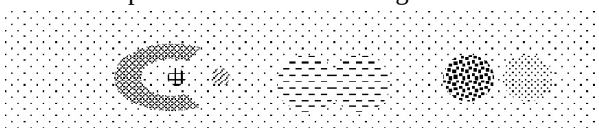


Figure 5: Six dense clusters of 2 dimensional points.

5) Similarity-based Cluster :A cluster is a set of objects that are “similar”, and objects in other clusters are not “similar.” A variation on this is to define a cluster as a set of points that together create a region with a uniform local property, e.g., density or shape.

3.VARIOUS PROBLEMS, ISSUES AND CHALLENGES IN CLUSTER ANALYSIS

3.1) Problems

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

1)The identification of distance measure :distance measures can be used for numerical attribute but for categorical attributes identification of distance measures is difficult.

2) The number of clusters:Identification of the number of clusters is a difficult task if the number of class labels is not known in prior.In order to produce correct results, a careful analysis of number of clusters is required.

3)Structure of database: In real life data, ,the clustersmay not always clearly identifiable . So the ordering of tuples and distance measures may affect the results when an algorithm is executed. With a structure less data (for eg. Having lots of missing values), identification of appropriate number of clusters will not yield good results.

4)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.

5)Classification ofClustering Algorithm: Clustering algorithms can be classified according to the method adopted to define the individual clusters[4][7].

3.2)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. There are following fundamental challenges which are associated with clustering The researcher must understand the following basic definition before proceeding to research.

What is a cluster?

What features should be used for cluster?

Is the normalisation of data is required?

How do we define the pair-wise similarity?

Is data consist of any clustering tendency?

Are the all clustering result easily visualized? Cluster Visualization is used for transformations from the problem domain” to the “representation domain”. Visualization is the critical challenge for clustering.

3.3)Issues in Cluster Analysis

We have grouped major issues of cluster analysis into following broad categories [9][13].Issue relared to Data , cluster, clustering algorithm, Data transformation, Cluster solution, Variable selection and Cluster validity are described in Table.1

ISSN: 2321-9653

Issues	Description
DataCharacterstics	• High dimensionality • Size of data • Sparseness • Noise and Outlier • Dataset and types of attributes • Scaling • Mathematical properties of data space
ClusterCharacterstics	• Distribution of data • Shapeof cluster • Different size of cluster • Differing densities of cluster • Poorly separated clusters • Relationship among clusters • Subspace clusters
Clustering method or technique	• Order dependence • Nondeterministic • Scalability • Selection of different parameters • Transforming the clustering problem to another domain
Data transformation	• 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	• 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 Of cluster	• 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 of 3?
Variable selection	• What is the best set of variables for generating a cluster analytic solution?

Table 1 Issues of Cluster

4.EVALUATION OF CLUSTERS

One of the main problem of clustering is to decide which number is an optimal number of clusters so that it fits considered data set. In order to evaluate how good results obtained after a clustering algorithm,a clustering optimal criterion is required. So the term cluater validity is used for evaluating the results of a clustering algorithm. Almost every clustering algorithm has dependence on the input parameters and features of the dataset .Incorrect input parameters may lead to clusters that deviate from those in the dataset. In order to determine those input parameters which lead to best fit clusters in a given dataset, we have required reliable guidelines to evaluate the clusters; clustering validity indexes have been recently employed. In general,compactness and separability are the terms which are used for defining cluster validity

Compactness: This measures how closely related objects in clusterare.
Separability: This indicates how distinct or well separated two clusters are.

ISSN: 2321-9653

The distance between representative objects of two clusters is a good example. This measure has been widely used due to its computational efficiency and effectiveness for hypersphere-shaped clusters.

There are three approaches to study cluster validity .[5][6][8]. The first is based on external criteria. This implies that the results of a clustering algorithm is evaluated on the basis of predefined knowledge which is imposed on a dataset.foreg. Class labels or no of clusters.The second approach is based on internal criteria. In this the results of a clustering algorithm is evaluated in terms of quantities and features that involves the vectors of the datasets themselves. Internal criteria can roughly be subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution. The third approach of clustering validity is based on relative criteria, which consists of evaluating the results (clustering structure) by comparing them with other clustering schemes.

Table 2. Evaluation criteria measures

Validity indices	Description
I.Internal Quality Criteria	Based on statistical methods
Sum of squared error (SSE)	SSE is suitable for that cluster which consist compactness and are well separated
Other minimum variance criteria	In this additional minimum criteria to SSE may be produced, they are also appropriate forwell separated clusters
Scatter criteria	It is obtained fromscatter matrices,It is used to reflect total scatter matrix which compute within and between cluster
Condorcet Criterion	Maximal number of clusters is not predetermined, it can be determined
C-Criterion	An extension of Condorcet criterion is called C-Criterion
Category utility metric	For small number of nominal features that have a small number of elements
Edge cut metrics	It is used for measuring quality as ratio of remaining edge weights to the total precut edge weghts.
Validating hierarchy clustering scheme (Cophenetic correlation criteria)	This index measure degree of similarit and used for validation of the clusters hierarchy
Validating a single clustering scheme(Hubert's ٢statistic (or normalized ٢ statistic)	This index find the degree of agrrement between aclustering scheme and proximity matrix.
II.External Quality Criteria	Based on statistical methods
Mutual information based measure	It can be used as an external measure for validity of clusters
Precision-recall measure	It is used for Verification of correctly clustered objects
Rand index	It is used to Compares an induced cluster with a given cluster.This index has value between 0 and 1
Monte Carlo techniques	This technique is used for Computation of probability density function for validityIndices.
Comparison of cluster structure with partition P	It is not useful for hierarchy of clusters.

ISSN: 2321-9653

Comparison of proximity matrix with partition P	It is mainly used for Comparisionof similarity between two matrices.
III.Relative Quality Criteria	This validation method is not based on statistical testing.It is based on relative criteria and choose best clustering scheme according toan a priori criterion. Validity indices for relative criteria are fuzzy clustering and crispClustering.

5.TOOLS

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 [9][13].

- Weka: Weka is a collection ofviaualisation tool and machine learning algorithms for data mining tasks and it is used for data analysis and predictive modeling.
- Matlab Statistical Toolbox : It is a collection of tools built on the MATLAB for organizing, analyzing and modeling .
- Octave: It is a free software similar to Matlab.
- CLUTO: It is software package for clustering low and high dimensional dataset.
- Databionic ESOM Tools : It is atoolthat offer many data mining tasks using emergent self-organizing maps (ESOM). Visualization, clustering, and classification of high-dimensional data using databionic principles can be performed interactively or automatically.
- SPAETH2 :It is a tool which is collection of routines for analysing data by grouping them into clusters
- XLMiner: It is a Professional level tool for data visualization,,forcasting and data mining in excel.It consist of machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction.
- DTREG: It is a statistical analysis program used for predictive modelling based on decision trees, SVM, Neural N/W and Gene Expression programs.
- Cluster3: It is an open sourceclustering software which provide graphical user interface to access cluster routines that can be used to analyze gene expression data.

6.CONCLUSION

Now a days clustering has become a subject of wide research since it arises in many application domains.Organizing data into sensible groupings arises naturally in many scientific fields. It is, therefore, not surprising to see the continued popularity of data clustering. It is important to remember that cluster analysis is an exploratory tool; the output of clustering algorithms only suggest hypotheses. In this paper we have covered the methods and types of clusters for cluster analysis . We have also discussed vaious problems, challenges and issues found in implementation and those which affect the clustering results.Another important issue that wehave described in this paper is the cluster validity. This is related to the inherent features of the data set under concern. At last we have described some of the software available that can ease the task of implementation.

REFERENCES

[1]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.

[2]Manish Verma, MaulySrivastava, NehaChack,"A Comparative Study of Various Clustering Algorithms in Data Mining"International Journal of Engineering Research and Applications (IJERA)Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384

[3]Maria Khalkidi,YannisBatistakis,Vazirgiannis "On Clustering Validation Techniques" Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001

[4]RuiXu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE "Survey of Clustering Algorithms" IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005

[5] Elena Sivogolovko "Validating cluster structures in Data Mining tasks" Proceedings of the 2012 EDBT/ICDT Workshops March 26-30, 2012, Berlin, Germany.

[6] EréndiraRendón, ItzelAbundez, Alejandra Arizmendi and Elvia M. Quiroz. "Internal versus External cluster validation indexes" International Journal Of Computers And Communication, Issue 1, Volume 5, 2011

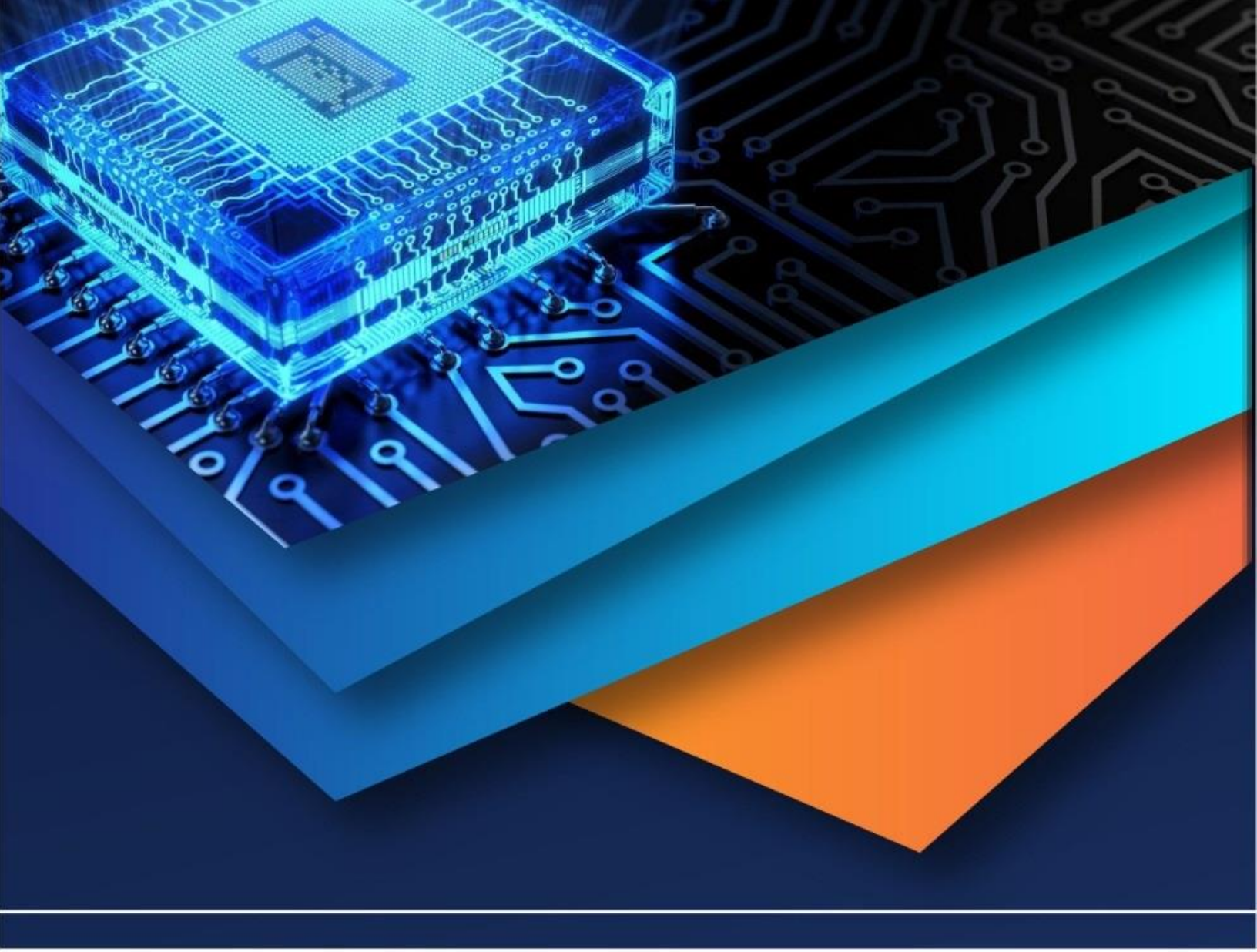
[7]Osama abuabbas" Comparison between Data Clustering Algoritm" The International Arab Journal of Information Technology , Vol.5, No 3, July 2008

[8]Raluca-Mariana Ştefan "Clustering Data For Knowledge" International Journal of Education and Research" Vol. 1 No. 5 May 2013ISSN: 2201-6333 (Print) ISSN: 2201-6740 (Online)

[9]V.Ilango,Dr.R.Subramanian "Cluster Analysis Research Design model, problems, issues, challenges, trends and tools", International Journal on Computer Science and Engineering (IJCSE)ISSN : 0975-3397 Vol. 3 , 2011.

[10] A.K.JAIN, Michigan State University,M.N. MURTY, Indian Institute of Science AND P.J. FLYNN, the Ohio State UniversityData Clustering: A Review, ACM Computing Surveys, Vol.31, No. 3, September 1999

[11]AmandeepKaur Mann "Survey Paper on Clustering Techniques" International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013, ISSN: 2278 – 7798



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY